Doctoral Thesis Proposal

Computational Methods for Inferring Biological Heterogeneity from Single-Cell Data

Sitara Persad
sitara.persad@columbia.edu

Advisor: Prof. Itsik Pe'er
Department of Computer Science
Columbia University

November 14, 2022

Abstract

Heterogeneity is a key feature of all biological systems, arising naturally from the evolution of complex systems from single-cell origins. From studying mechanisms that drive well-ordered development in healthy systems, to identifying tumour heterogeneity and mechanisms of tumour resistance to therapeutic intervention, quantifying and tracing the roots of biological heterogeneity has important biological significance across many areas.

Traditionally, bulk sequencing of genetic material has been used to gain insight into biological systems; however, such population averages often present a misleading picture of the true programs at work in a heterogeneous population. Recent advances in single-cell sequencing have drastically increased the potential for investigating heterogeneous systems at high resolution, but such experimental methods come with their own limitations in the form of noise and computational complexity. Furthermore, such methods only provide a snapshot of the system at present, and inference must be performed in order to gain insight into the processes that drive heterogeneity.

In this thesis proposal, we discuss current challenges in understanding biological heterogeneity from single-cell sequencing, and present our work aimed at addressing these challenges. We present SEACells, our method for addressing noise and sparsity in single cell data, in order to enable more robust downstream inference of mechanisms that drive heterogeneity. Our method partitions cells into highly similar cell states using kernel archetype analysis, and aggregates them into more robust entities. We also discuss our ongoing work on analysing heterogeneity in abnormal development, by inferring tumour intra-heterogeneity from copy number variation using Hidden Markov Models. Finally, we propose an algorithm for reconstructing lineages from CRISPR-Cas9 lineage recording single-cell data, in order to study the origin and mechanisms of heterogeneity in a wide range of biological systems.

Contents

1. Introduction

The development of sequencing methods allowed scientists to determine the order of the nucleic acids in biological samples, and thus decode the sentences written in the four character alphabet of **A**denine, **C**ytosine, **G**uanine and **T**hymine that constitute the instructions and playbook for life. With the advent of massively parallel sequencing, the cost of sequencing has dramatically decreased, while the throughput has increased by orders of magnitude [64], enabling the widespread adoption of such sequencing technologies. Until recently, these sequencing methods all involved an initial step of pooling the genetic material of thousands or millions of cells before sequencing, thus earning the technique the title *bulk sequencing*. This has an obvious downside - for example, bulk transcriptomic studies which measure gene expression levels observe the average behaviour of the populations under investigation. Such analyses assume homogeneity amongst all cells in the sample, or require complex deconvolution procedures to attempt to distinguish between distinct populations present in the sample.

The characterization of biological samples by their population average can be grossly misleading, as heterogeneity is a key feature in most biological systems [27,2,21,31]. From studies of epigenetic programs that drive well-ordered development and differentiation in healthy cells to the formation of distinct abnormal subpopulations of cancerous cells in a tumour, heterogeneity is a hallmark of biological systems, and the ability to identify and study distinct phenotypes in heterogeneous populations is paramount. Beyond gaining new insights into the function of tissues and organs, studying heterogeneity in diseased tissues can allow for the development of new therapeutic approaches or allow us to understand and characterise why treatments may fail [29].

The origin of heterogeneity is often considered in genetic terms; generated by random mutations which evolution acts upon via natural selection. In such models, we can trace the genetic differences between populations to a lineage tree that describes the relationships between distinct populations. However, it is now understood that the sources of biological heterogeneity are multifaceted [1,28,15]. Epigenetic control, such as chromatin accessibility, nucleosome positioning, histone tail modifications and enhancer–promoter interactions, is another important driver of heterogeneity. In order to gain a full understanding of biological heterogeneity, we must therefore consider both genetic and epigenetic variation.

Within the past decade, single-cell sequencing approaches for probing the genome and epi-genome have been developed and risen in popularity [46,33]. In these methods, individual cells are isolated and barcoded so that any data collected can be identified with a single cell. This allows scientists to solve the problem of averaging over heterogeneous populations and perform a wide range of analyses that were previously difficult to nearly impossible, such as characterising the similarities and differences between populations of cells, identifying rare cell types, or trace lineages and developmental relationships in systems such as embryonic development and cancer.

However, single-cell sequencing data is not without its own set of limitations, unique to each data modality. For the purpose of this work, we will limit our focus to the use of and challenges in single cell RNA-sequencing and single cell ATAC sequencing for inference of biological heterogeneity.

1.1. Proposal Structure

The organisation of this proposal proceeds as follows. In Section 2, we review previous research on extracting insights into biological heterogeneity from single cell transcriptome and epigenetic assays, particularly in the context of phylogenetic relationships which drive heterogeneity. We introduce and discuss current work on reconstructing cell lineages from induced CRISPR scars. Section 3 describes our completed work on overcoming single-cell RNA and ATAC noise and sparsity by aggregation into metacells. In Section 4, we discuss our ongoing work on inference of copy number alterations for investigating tumour heterogeneity. In Section 5, we discuss our proposal for CRISPR lineage reconstruction. Finally, in Section 6, we outline the research plan for completing this thesis.

2. Background and Related Works

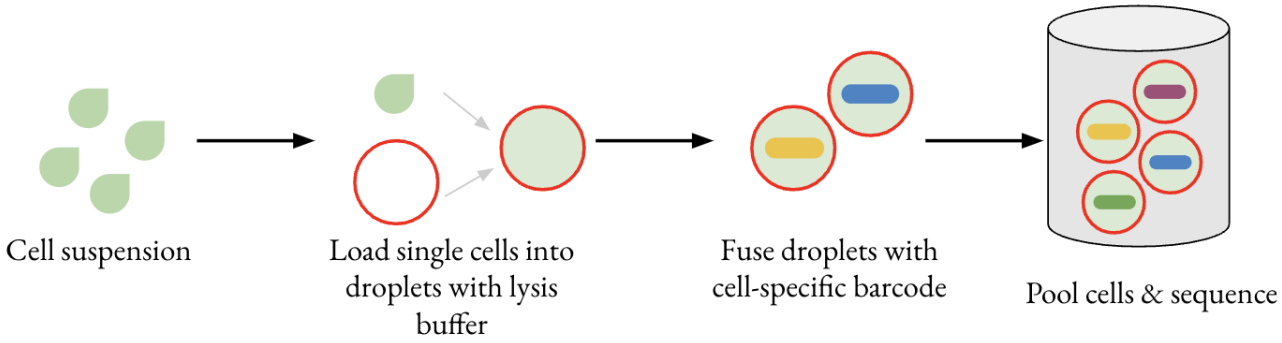
2.1. Challenges in single-cell data science

In 2013, single-cell sequencing was highlighted as the "Method of the Year"; advancements in microfluidics and combinatorial indexing strategies had enabled a data revolution in single-cell biology where experiments involving the collection of thousands to millions of single-cell measurements became routine[43]. In more recent years, single cell measurements of not only RNA, but also chromatin accessibility, has enabled us to investigate transcriptome-wide gene expression and epigenetics at a single-cell resolution for millions of cells. The ability to probe large scale systems at such resolution has drastically increased the potential for investigating heterogeneity, but not without accompanying novel challenges [38]. These challenges can be broadly divided into two categories:

(1) Dealing with noisy and sparse single-cell measurements
(2) Dealing with single-cell data at scale

2.1.1. Noise in single-cell RNA sequencing

The pipeline for generation of single-cell RNA sequencing follows the general outline described below, and illustrated in Figure 2.1. First, single cells are isolated from a cell suspension into droplets containing a lysis buffer. The lysis buffer is a solution which breaks open the isolated cell in order to analyse the mRNA molecules within it. After lysis, each droplet is fused with a cell-specific barcode as well as enzymes for reverse transcription. The cells, which can now be uniquely identified by their barcodes, are pooled and reverse transcription is triggered. Droplets are broken and the complementary DNA (cDNA) is purified. This cDNA is then further amplified before sequencing.



2.1.1.1. Figure 2.1. Experimental procedure for single cell RNA sequencing

However, there may be limited amounts of information per cell, which leads to high levels of uncertainty about observations in any given cell. Current scRNA-seq protocols generally capture a small random sample of 10-50% of the mRNAs present in a cell [73,62]. If we use increasing levels of amplification to generate more material, there is a corresponding increase in technical noise present in the data [38]. Technical sources of such sparsity can obscure the difference between true biological zeros, where a gene is not present in a cell, and methodological noise.

One approach to addressing the issue of sparsity is imputation, whereby a value for missing data is inferred. There are three main approaches used to perform imputation:

(1) Model-based imputation, where a probabilistic model is used to identify which observed zeros are likely to be technical rather than true zeros. [6,16,39]
(2) Data smoothing methods, which define similarity between a cell and its neighbours and adjust its gene expression values based on values in similar cells [69,65,26]
(3) Data reconstruction methods which define a latent space representation of the cell and reconstruct the observed data matrix from the simplified representation [40,47,5]

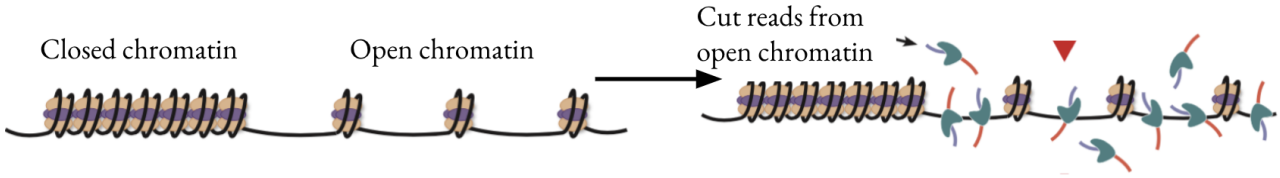
However, imputation methods suffer from several drawbacks. Relying on information solely within the imputed dataset may lead to artificial amplified signals, which can inflate correlations between genes and cells and lead to false positives in downstream analyses [3]. Another drawback of relying on latent representations for reconstruction of gene expression values is that such imputed counts are often not trusted by biologists, as they do not represent 'true' data points which were observed in the experiment.

In summary, current methods for overcoming sparsity and technical noise in scRNA-seq suffer from the problems of inflating correlations in the dataset, having to walk a fine line in balancing between under- and over-correcting technical noise between experiments and relying on latent spaces to generate simulated gene expression counts which may not be trusted by biologists.

2.1.2. Noise in single-cell ATAC sequencing

The adaptation of the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) to single-cell sequencing is the state of the art method for analysis of the genome-wide

chromatin accessibility landscape at single cell resolution [12]. As illustrated in Figure 2.2, single-cell ATAC-seq allows us to interrogate open chromatin regions, which are special regions of the human genome that can be accessed by DNA regulatory elements. Transcriptional activity has been tightly linked with disruption to DNA packaging at promoter, enhancer, silencer and other locus control regions. Open chromatin regions therefore coincide with regulatory DNA, so investigating patterns of chromatin accessibility provides us with a window into the epigenetic regulation of gene expression [63].



2.1.2.1. Figure 2.2 scATAC-sequencing allows us to probe open chromatin regions by sequencing accessible DNA

The problem of sparsity is even more acute in scATAC-seq, as the number of potential cis-regulatory elements in a cell far exceeds the number of sequence reads. To illustrate, a standard scATAC-seq experiment contains 10^3-10^5 sequencing reads, but hundreds of thousands of loci of accessible chromatin . Furthermore, even for a regulatory element which has been covered with reads, the number of mapped reads rarely exceeds two, since each locus has no more than two copies of a region of chromatin per cell in a diploid cell [36]. As such, the dynamic range of scATAC-seq is far lower than scRNA-seq, where there may be many more reads corresponding to expressed genes.

Existing methods attempt to remedy the issue of sparsity via two main strategies:

(1) Aggregate reads across regulatory elements - instead of analysing each regulatory element individually, these methods combine regions that share either a transcription factor binding motif or some co-activation pattern. They then aggregate peak counts across regions to cluster cells or perform other downstream analysis [36,56,35]. However, any information specific to a regulatory element is lost, and such methods rely on known motifs or co-activation patterns. Furthermore, such pooling may still be insufficient to overcome sparsity concerns.

(2) Pooling cells across cell types - instead of analysing each cell individually, these methods compute a pseudo-bulk sample by aggregating cells of a similar cell type [49,79]. Pooling across cell-types may result in loss of heterogeneity, as signals are likely averaged across distinct cell states.

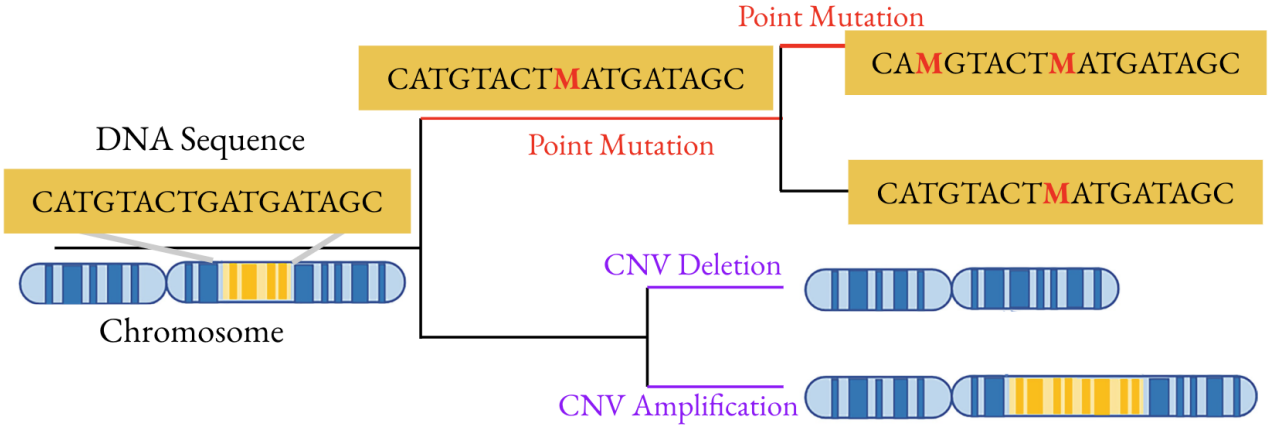
2.1.3. Scaling to Large Single-Cell Datasets

The scale of single-cell datasets has been rapidly expanding, as single-cell throughput has increased, experiments have become cheaper and more widely adopted and consortium efforts have led to massive aggregated datasets [4]. The volume of data will soon become a challenge for existing pipelines, further motivating the need for computational approaches that can make such volumes of data tractable. Existing methods include subsampling approaches, which reduce dataset size at the cost of potentially missing out on rare populations. Structure-aware subsampling methods can mitigate the risk of such losses [34], but still do not take full advantage of large dataset size.

In order to extract full value from increasing quantities of (noisy and sparse) single cell sequencing data, there remains a pressing need for methods that allow us to correct for these challenges, in order to enable downstream analysis of biological heterogeneity. Aggregation of samples into granular cell states is a promising approach to combating this issue. In Section 3, we describe our work on addressing these challenges for single cell RNA and ATAC sequencing.

2.2. Inferring Intra-Tumour Heterogeneity

Cancer is a dynamic disease, and tumours evolve over the course of disease progression, increasingly acquiring genetic heterogeneity over time [18]. A tumour is not a uniform mass, but rather contains a non-uniform distribution of phylogenetically related subclones, each of which has independently acquired distinct somatic mutations. This phenomenon is referred to as intra-tumour heterogeneity. The somatic mutations which differentiate subclones within a tumour may take the form of both point mutations and copy number alterations, as illustrated in Figure 2.3. Copy number alterations occur when regions of chromosome, which are usually present in sets of two in normal cells, are amplified or deleted. In such a case, a cell is no longer diploid, meaning that they contain two copies of each gene, but may instead contain variable numbers of copies of a given gene in a copy number altered chromosomal region.



2.2.0.1. Figure 2.3. Tumours evolve through the accumulation of genetic alterations, such as point mutations and copy number alterations, whereby regions of the chromosome may be duplicated or deleted

2.2.1. Reconstructing Tumour Phylogeny

As intra-tumour heterogeneity is an evolutionary process, heterogeneous subclones are related to each other by a phylogenetic tree (Figure 2.3). Access to this phylogeny allows us to move beyond simply identifying heterogeneous populations, but also gain an understanding of the evolutionary processes involved in cancer, such as order and timing of driver mutations, processes which drive metastasis or resistance to therapy. Currently, most existing methods for inferring tumour phylogeny rely on constructing trees from point mutations using single cell DNA sequencing [77,51]. Unfortunately, scDNA-seq remains quite costly and datasets consist of relatively few cells compared to other high throughput methods. Furthermore, the use of scDNA-sequencing rather than scRNA-sequencing means

that there is no access to transcriptional information, which describes functional behaviour of cells. Finally, these methods which focus on solely point mutations neglect the impact of copy number alterations altogether, leaving this as an understudied field in tumour phylogenetics.

2.2.2. Copy Number Alterations from scRNA-sequencing

Copy number instability is a hallmark of cancer [32], and represents a significant source of intra-heterogeneity [70,68]. Despite this fact, most existing studies of intra-tumour heterogeneity have focused on somatic mutations, but further attention must be paid to identifying heterogeneity in copy number alterations, which have been shown to be a strong predictor of survival rate across cancers [67]. In order to use copy number alterations to characterise intra-tumour heterogeneity, we must first be able to infer the presence of aberrations in copy number in single cell RNA sequencing.

There are currently three major classes of methods used for inference of copy number alterations:

(1) Sliding window approach - InferCNV [45] uses a moving average of gene expression data to determine CNV profiles.
(2) Hidden Markov Model - CaSpER [57] and HoneyBADGER [19] use a Hidden Markov Model on gene expression and minor allele frequencies to infer the presence of copy number deletion or amplification.
(3) Markov Chain Monte Carlo - CopyKAT [22] infers a diploid reference from hierarchical clustering and computes relative gene expression for predicted tumour cells. It then uses a Poisson-gamma model and Markov Chain Monte Carlo to generate a posterior mean per genomic window, to infer the position of breakpoints where copy number changes across the genome.

These approaches rely on gene expression as a proxy for copy number alteration, but none take into account the fact that different cell types also influence gene expression levels. Furthermore, they do not learn or make use of the phylogenetic relationship between clones, where clonal relationships can be used to inform prediction of copy number alterations and vice versa. In Section 4 of this thesis proposal, we discuss our work on PICASSO, a phylogenetically driven model for inferring copy number alterations from scRNA-seq, whilst addressing these limitations.

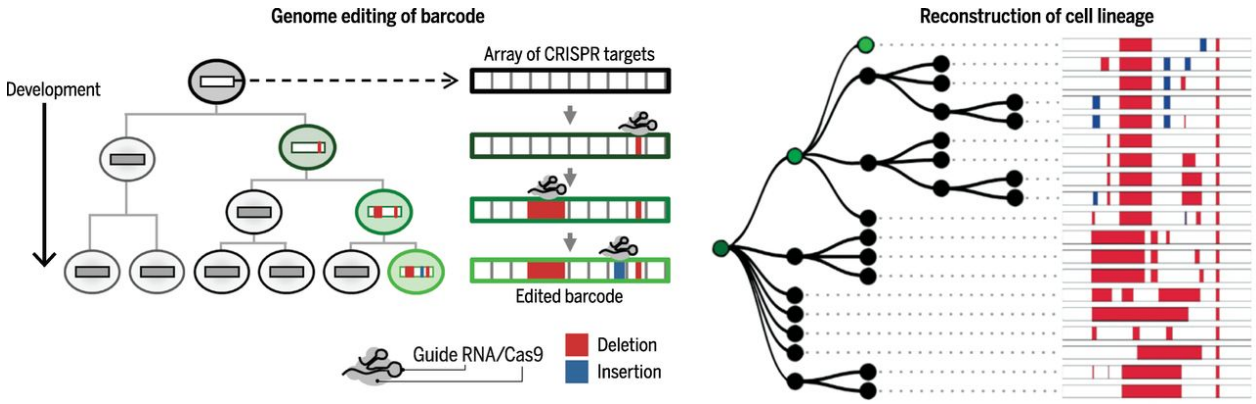
2.3. Lineage Tracing via CRISPR-Induced Scarring

Constructing tumour phylogenies from copy number alterations is a noisy and challenging process due to the presence of recurrent copy number alterations across subclones, and inherent challenges in inferring copy number alterations. Furthermore, beyond the abnormal evolution of cancer which contains mutations that allow us to gain insight into the lineage of that system, normal developmental processes do not contain such information. Lineage tracing refers to the set of methods which allow us to reconstruct the cell fate decisions that lead to the formation of complex cell systems with heterogeneous cell states, with minimal perturbation to the system itself [74]. Recently, methods using genetic labelling have enabled the permanent tracing of cells' descendants. The development of CRISPR-Cas9 genome editing,

in conjunction with single-cell RNA-sequencing, now provides us with a tool for conducting large scale lineage tracing at the single cell level while simultaneously reporting transcriptomic data [76].

2.3.1. CRISPR-Cas9 Editing Creates Heritable Alterations in DNA

Whilst the number of experimental approaches for collecting lineage tracing data is ever-growing [42, 61, 60, 55], the general experimental setup for CRISPR-Cas9 editing systems follows the following pattern (as illustrated in Figure 2.4). A synthetic construct containing multiple target sites is inserted into the genetic material of the founder population of cells. Upon induction of CRISPR-Cas9 activity, this construct accumulates stochastic insertions and deletions (indels) at each target site. These indels are inherited by any offspring of that cell. Later, the offspring cells are sequenced and the patterns of inherited indels can be used to trace their evolutionary history via a phylogenetic reconstruction algorithm[24].



2.3.1.1. Figure 2.4. CRISPR-Cas9 lineage tracing experimental setup. A construct containing CRISPR targets records insertions and deletions which are ultimately sequenced and used to reconstruct lineage [42]

2.3.2. Reconstructing Lineages from CRISPR Edits

As CRISPR-Cas9 lineage tracing is a relatively recent experimental method, there is not yet a widespread consensus of the ideal algorithm for phylogenetic reconstruction. Methods from classical phylogenetics which either aim to construct a tree by minimising distances between neighbouring leaves [54], or by constructing a maximum-parsimony tree which requires fewest mutations to explain the data [13], have been employed for reconstructing lineages from CRISPR-Cas9-induced insertions/deletions [42]. However, such lineage recording data have different noise characteristics and evolutionary models from classical mutation datasets, and, as such, these traditional phylogenetic algorithms may not be well suited to large-scale lineage tracing experiments.

The DREAM challenge for lineage reconstruction from CRISPR recorder systems [24] benchmarked five top performing approaches for lineage reconstruction, which fell into two main categories:

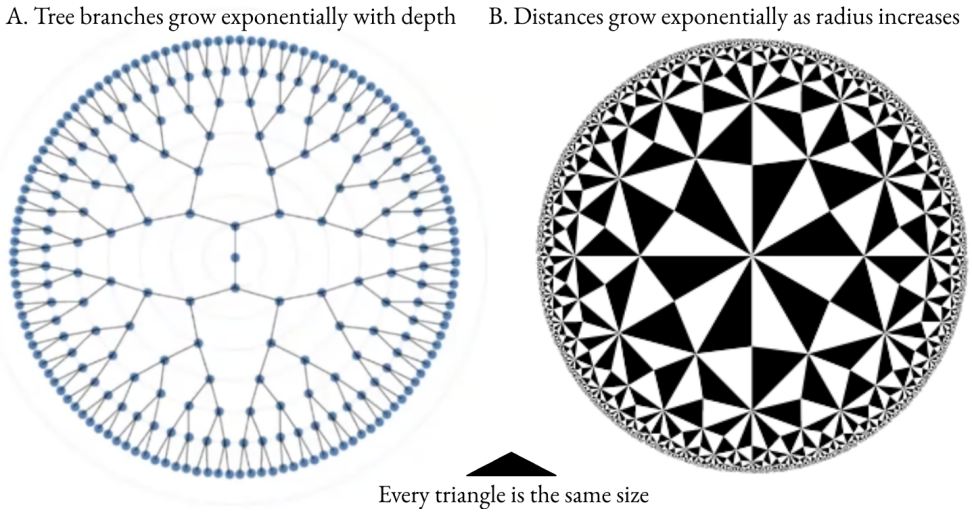
(1) Distance-based methods - Cell-cell distance matrices serve as input to hierarchical clustering methods that give the resulting tree. Distance matrices were adapted to CRISPR-Cas9 systems by

taking into account non-uniform distribution in indel probabilities or unidirectional mutation from unedited to edited state [25].

(2) A combined greedy/combinatorial optimization approach which splits the tree greedily based on mutation frequencies and then infers a Steiner tree for subtrees[37]. In the case of independently recurring mutations, however, such greedy top down approaches will construct incorrect trees.

2.3.3. Hyperbolic Optimization for Tree Reconstruction

Recent work in artificial intelligence has shown that using hyperbolic space to represent latent hierarchies yields significant performance improvements over traditional methods [44]. Hyperbolic geometry is particularly well suited for tree optimization as it can more accurately represent a hierarchical structure compared to Euclidean space [41]. This is because as tree depth increases, the number of leaves increases exponentially, leading to overcrowding in Euclidean space. In hyperbolic spaces, the distance grows exponentially with radius, just as the number of nodes grows exponentially with tree depth (Figure 2.5.)



2.3.3.1. Figure 2.5. Hyperbolic spaces such as Poincaré balls (B) are well suited to modelling hierarchical data (A) as distances between points grows exponentially

If we optimise distances between leaves in hyperbolic space we can get arbitrarily close to a true tree metric, which can be used to compute a more accurate phylogeny. Recent work has used Riemannian optimization to optimise phylogenetic tree distances according to a log-likelihood model defined by mutation probabilities [71]. Such methods are well suited to large scale data, as they perform gradient descent in a continuous space rather than discrete optimisation of tree configurations which may lead to combinatorial explosion. In Section 5, we discuss our proposed work on reconstructing lineage trees using hyperbolic optimisation.

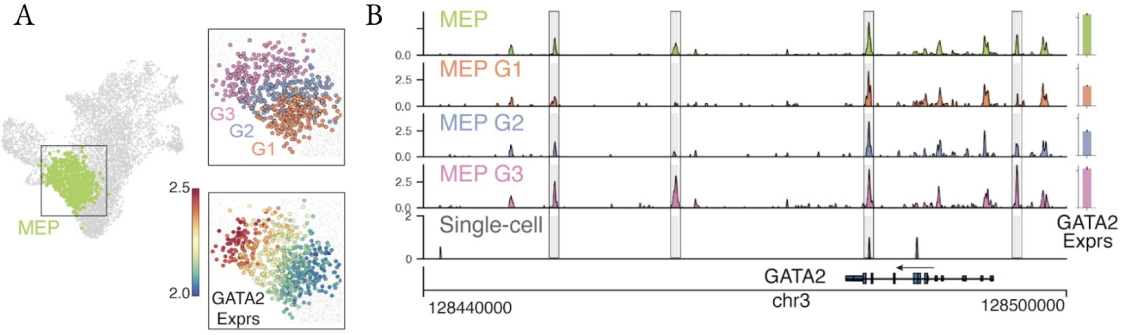
3. Kernel Archetypal Analysis for scRNA- and scATAC-seq Metacells

3.1. Introduction

A fundamental disconnect currently exists between the cellular resolution of single-cell genomics data and the cluster-level resolution of analysis, which has dramatically limited these technologies in fulfilling their potential for biomedical research. A dataset that harbours tens of thousands of cells is typically summarised as a handful of clusters in order to overcome the noise and sparsity inherent to single-cell data. Sparsity is particularly acute in scATAC-seq data, which only captures the trinary zygosity states at a few thousand of the hundreds of thousands of open chromatin regions in a cell, making it impossible to infer regulation at the single-cell level. While scRNA-seq data is not as sparse, projects such as the Human Cell Atlas [53] and Human Tumour Atlas Network [52] are scaling to millions of cells, causing even routine dimensionality reduction and visualisation tasks to struggle with computational complexity. As a result, large scRNA-seq datasets are also typically analysed at the cluster level.

3.2. Limitations of Previous Approaches

Cluster-level analysis has led to important biological discoveries. However, a typical cluster is not homogenous. Moreover, single-cell data has been shown to reside on a continuum. Such dynamics are lost in any discrete cluster-level analysis (Figure 3.1).



3.2.0.1. Figure 3.1. A. UMAP with megakaryocyte-erythroid progenitor (MEP) cluster highlighted, then split into three equal-sized bins based on developmental progression (top), reflecting imputed expression of GATA2 (known driver of MEP lineage) (bottom). B. Coverage plots showing GATA2 accessibility in all MEPs (top), a single MEP cell (bottom) and in the three bins in (C). Accessibility dynamics track with expression dynamics, but is masked at cluster level, whereas peak identification in single cells is too noisy.

The concept of metacells [7] - groups of cells that represent distinct, highly granular cell states, whereby within-metacell variation is due to technical rather than biological sources - was proposed as a way to maintain statistical utility while maximising effective data resolution. Metacells are far more granular than clusters, and are optimised for homogeneity within cell groups, rather than for separation between

clusters. However, existing approaches [7, 10, 9] fail on scATAC-seq data, aggressively cull outliers (particularly inappropriate for disease studies, which are often driven by rare cell populations), and are poorly distributed across the phenotypic space. Consequently, metacells have not been routinely used in single-cell analysis, and scATAC-seq data has remained heavily underutilised.

To address these limitations, we developed SEACells,- **S**ingle-c**E**ll **A**ggregation of High-Resolution **Cell-s**tates - a graph-based algorithm that uses kernel archetypal analysis on the nearest neighbour graph to compute metacells in both scRNA-seq and scATAC-seq which capture rare cell-types and recover cell states spanning the phenotypic manifold.

3.3. Method

The SEACells algorithm assumes that biological systems consist of well-defined and finite sets of cell-states defined by co-varying patterns of gene expression. Observed single-cell data are assumed to be noisy measurements of these cell-states with current state of the art single-cell measurement technologies able to capture <10% of transcripts or <5% open chromatin regions. Despite the high degree of noise, cells sampled from the same states are assumed to be closely related in their phenotypes as a result of gene expression patterns and regulatory mechanisms that define the cell-states. Thus, SEACells algorithm aims to aggregate closely related single-cells into metacells representing them. As a result of aggregation, metacells overcome the sparsity that plague single-cell data, with single-cell ATAC-seq data particularly limited in its utility due to its sparsity. SEACells metacells also provide a scalable representation to efficiently handle large-scale single-cell data.

The inputs to SEACells algorithm are: (i) raw count matrices (E.g.: gene expression for RNA, peak or bin counts for ATAC), (ii) low dimensional representation of the data derived using an appropriate preprocessing procedure dependent on data modality such as PCA for RNA and (iii) the number of metacells to be identified. As output, SEACells produces groupings of cells that represent metacells, aggregated metacells X feature raw counts matrices and soft assignments representing highly related groups of cells, all of which can be used for downstream analysis.

SEACells comprises five main steps:

(1) Construct a k-nearest neighbour graph using Euclidean distances between cells, computed in the lower dimensional embedded space,which provides a representation of the phenotypic manifold.
(2) An affinity matrix of cell-to-cell similarities is derived using the nearest neighbour graph. The affinity or kernel matrix encodes the non-linear relationships between cells .
(3) The kernel matrix then serves as input to kernel archetypal analysis. Kernel AA decomposes the data into an archetype matrix, linear combinations of cells that are representative of the cell-states on the phenotypic manifold and a membership matrix that reconstructs the single-cells as linear combinations of archetypes
(4) The groupings identified through archetypal analysis are SEACells metacells. Single-cell raw counts are aggregated using these groupings to derive a metacell X feature count matrix.

(5) Normalised metacell count matrices can be used for all downstream tasks including clustering, visualisation, integration, trajectory inference, ATAC-seq based regulatory inference and more.

3.3.1. Kernel Matrix Construction

The input to the kernel archetype analysis algorithm is a kernel matrix constructed from a k-nearest neighbour graph. This graph is constructed using Euclidean distance in the low-dimensional embedding (PCA or SVD). These distances are transformed into cell-cell similarities using an adaptive Gaussian kernel [66], which corrects for the remarkable variability in data densities present in single cell data, particularly in regions of rare cell-types which are of the most biological interest.

In this kernel space, two cells (x and y) are embedded close to each other if they satisfy two conditions:

1. x and y share neighbours in the PCA/SVD space
2. the similarity score among the neighbours of x and y are similar

Two cells in this transformed dimensional space will be similar to each other not only if they share the neighbours but only if the distances to the shared neighbours are also similar, imposing stricter similarity conditions between cells. The full details of kernel matrix construction are described in Appendix A.

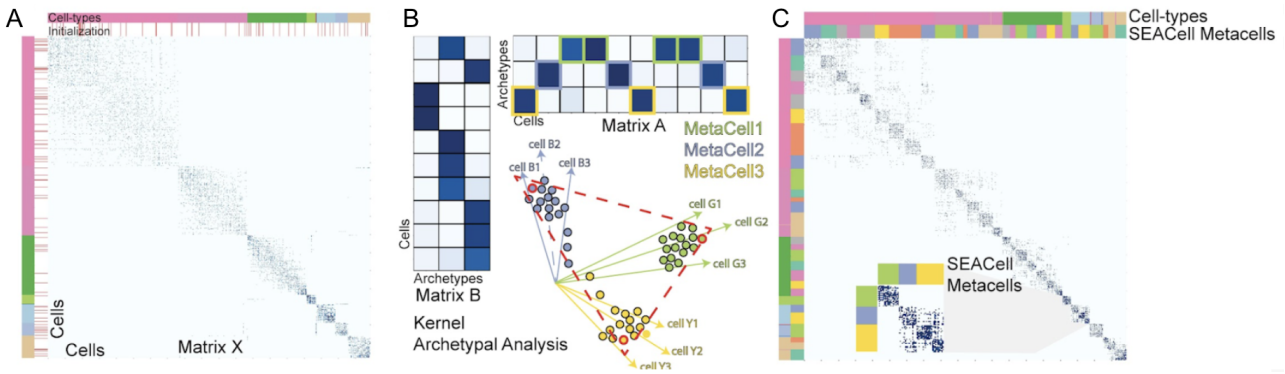
3.3.2. Kernel Archetype Analysis

Overview and optimization function

The adaptive Gaussian kernel matrix, $M \in R^{n\,X\,n}$, serves as input to an archetypal analysis procedure. Archetypal analysis [17] performs a linear decomposition of the kernel matrix where the goal is to identify a specified number of archetypes that are each a linear combination of the data points represented by the archetype matrix, $B \in R^{n \times s}$. The data points themselves are represented as a linear combination of the archetypes in a membership matrix, $A \in R^{s \times n}$, to reconstruct the kernel matrix. The number of archetypes, s, is substantially lower than the number of data points and the lower dimensionality of the archetype and membership matrices creates an information bottleneck, ensuring an optimal decomposition of the data [8]. The membership matrix contains weighted assignments of cells to archetypes, and can be used to derive cell partitions that are aggregated to metacells. This process is illustrated in Figure 3.2 below.

Taken together, the objective of archetypal analysis is to find matrices A, B such that product MBA forms a faithful reconstruction of the original kernel matrix by minimising squared reconstruction error:

$$\min_{A,B} SRE = ||M - MBA||^2 = tr[M^T M - 2M^T MBA + A^T B^T M^T M]$$



3.3.2.1. Figure 3.2. A. Kernel matrix serves as input to archetype analysis. B. Kernel AA identifies highly connected cells as archetypes, C. partitioning cells into groups of highly related cells

Archetype analysis in this kernel space is well suited to identify metacells. Kernel space is a Cartesian space, where the "phenotype" of a cell is defined by its similarity—quantified in the kernel matrix (M)—to every other cell. By definition, similarity scores are between 0 and 1; thus, all cells are in the positive quadrant, encapsulated by a hypersphere of radius strictly less than $\sqrt{N}$, where N is the number of cells. Intuitively, the kernel trick is casting highly similar cells into tiny clusters along a cone emanating from the origin. Within each small group of highly similar cells, the most representative cell of the group has to be most similar to every other cell in that group. In other words, the best representative of the unique cellular state of the highly similar cell group is the one that is most connected to every other cell in the group. Such a best candidate is likely to exist at the extreme (archetype) in the kernel space, as illustrated in Figure 3.2-B.

Optimization Algorithm for Metacell Identification

The objective function for kernel archetype analysis involves optimising the non-convex product AB, and thus has many local minima. The objective function is, however, convex in A given a fixed B matrix, and vice versa. Therefore, alternating minimization of weight matrices A and B is used to make the problem of solving archetypal analysis more tractable. The gradient function to update each matrix is a linear function in that matrix, whose optimal solution set lies on the vertex of the simplex defined by the matrix. Given this, we use Frank-Wolfe updates to optimise each weight matrix in turn.

Initialization

As archetypal analysis is a non-convex problem, solutions depend on the initialization of archetype and cell assignments. Given the density differences in the phenotypic manifold, random sampling of cells will lead to significant overrepresentation of initial points in the high-density regions and severe underrepresentation of cells in the biologically critical low-density regions. Therefore, we employ max-min sampling of waypoints to initialise archetypes. Given a set of waypoints, each additional waypoint is chosen to maximise the distance to the current set, i.e. maximise the minimum distance to any of the points in the current set. This ensures that the sampled waypoints are uniformly distributed across the phenotypic manifold irrespective of the density.

Summarisation by Aggregation
Archetypes are an approximation to the convex-hull i.e., archetypes represent the vertices of a convex polytope that encapsulates most of the data. By definition, archetypes are a linear combination of observed data points and hence do not necessarily represent measured data points themselves. Also by definition, each cell is expressed as a linear combination of the inferred archetypes. In order to aid in interpretability and facilitate downstream analysis, metacells are constructed by (1) computing binarized assignments of cells to archetypes (of the A matrix) and (2) aggregating single cells assigned to each SEACell by summing over raw counts. This summarised SEACells data matrix is significantly less sparse and noisy and can then be used for more robust downstream analysis.

3.4. Results

3.4.1. Metrics

Given that metacells represent distinct cell-states of a biological system, inferred metacells should be (i) compact i.e., low variability amongst cells that are aggregated with most of the variability a result of measurement noise and (ii) well separated from neighbouring metacells, since distinct metacells should include distinct gene-gene covariation matrices, even if these distinctions are subtle.

We used diffusion components to quantify both the compactness and separation of metacells. Diffusion maps have been used extensively to robustly and faithfully represent the phenotypic manifold using single-cell data [59, 30]. Each diffusion component represents a key axis of biological variance in both continuous trajectories and discrete states and thus provides an ideal platform to quantify metacell qualities.

Compactness
For each metacell, the variance in each diffusion component dimension is computed across its constituent cells. The average variance across components is reported as the compactness. Since diffusion components are by definition orthonormal, we can compute the variance of each component separately. The average variance ensures that the homogeneity of cells that constitute the metacell are measured across all axes of biological variance. A high quality metacell should have a low compactness score indicating homogeneity amongst the cells that constitute the metacell.

Separation
To assess whether metacells are distinct from each other, we evaluated the separation between neighbouring metacells using diffusion components. For each metacell, diffusion embedding is determined as the average of the cells that constitute the metacell. Distance between the metacell and its nearest neighbour is reported as the separation of the metacell. Agreater distance between metacells determined in diffusion space indicates a better separation between them.

Cell-type Purity

Cell-type purity is a measure of the consistency of cell-types amongst cells that constitute a metacell and was introduced to assess the quality of Super-cells [10]. Cell-type purity is computed as the proportion of cells which belong to the modal cell-type in a metacell. Note that purity metric is suited for biological systems that comprise distinct cell-types with distinct functions such as PBMCs, but less reliable for continuous trajectories such as CD34+.

3.4.2. Datasets

We perform benchmarking using three datasets:

(1) A CD34+ multiome dataset of 6881 cells containing simultaneous scRNA-seq and scATAC-seq for each cell during hematopoietic differentiation.

(2) A public 10X multiome dataset of peripheral blood mononuclear cells (PBMCs) [23] as a well-studied system with distinct cell populations. This dataset contains 11543 cells.

(3) Mouse gastrulation data of approximately 116,000 single cells across a range of cell types, containing scRNA-seq data for each cell during an early developmental process [48].

3.4.3. Benchmarking

We first visualise metacell assignments on two-dimensional projections on CD34+ single-cell ATAC sequencing data (Figure 3.3 below). We colour each cell by metacell assignment, and plot each metacells position as the average coordinates of its constituent cells. We can qualitatively observe on these low dimensional representations that SEACells preserves the structure of scATAC-seq much better, producing coherent metacells that span the phenotypic manifold.

3.4.3.1.

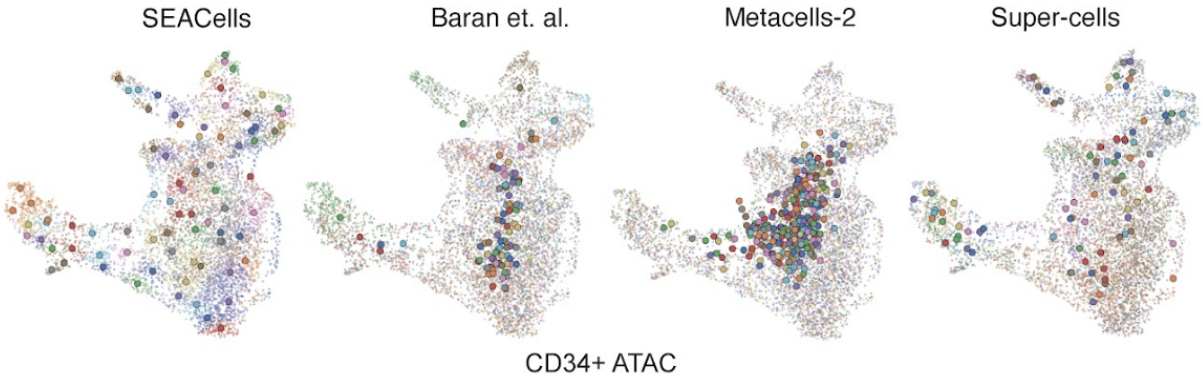


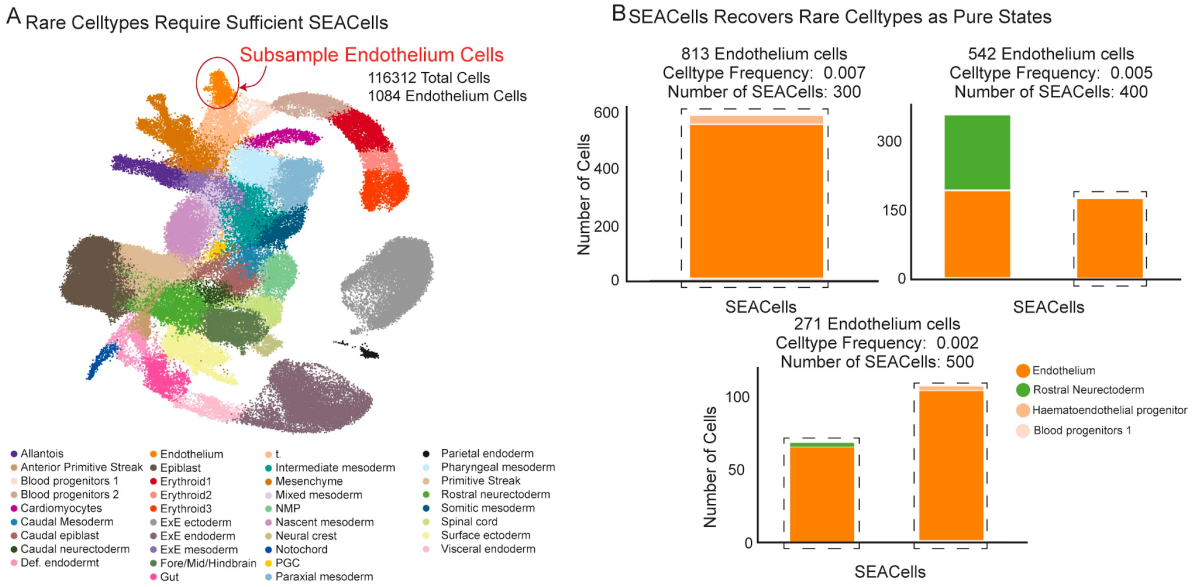
Figure 3.3. UMAP showing metacell groupings in CD34+ scATAC-seq across existing methods.

We quantitatively evaluated all methods for both continuous CD34+ and discrete PBMC datasets using the metrics described in the section above, shown in Appendix B.

3.4.4. SEACells Recovers Rare Cell-Types

To systematically assess the sensitivity of SEACells to capture rare cell states, we perform a downsampling experiment using the mouse gastrulation dataset. We subsampled different fractions of endothelial cells

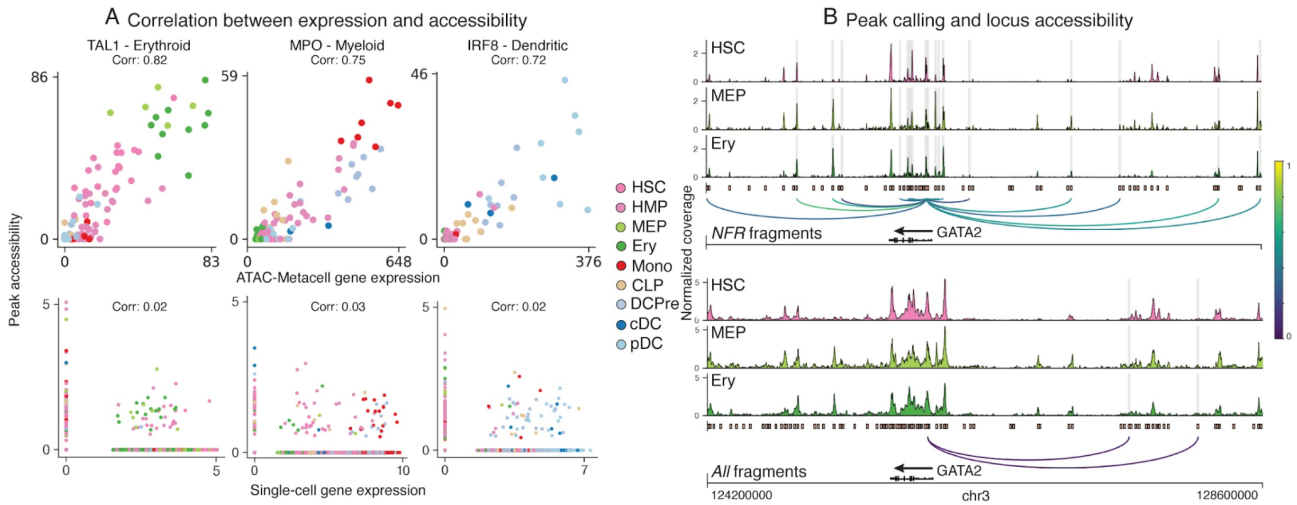
(0.7%, 0.5% and 0.2%) from the data while retaining all other cells and applied our SEACells algorithm to compute metacells. Following the application of SEACells, we examined all metacells in which endothelial cells constituted at least 50% of the cells that define that metacell and find that SEACells is able to recover metacells purely composed of endothelial cells, even at very low cell-type frequencies.



3.4.4.1. Figure 3.4. A. UMAP of the mouse gastrulation data with endothelial cells highlighted. B. Barplots showing the cell-type composition of metacells containing at least 50% endothelial cells with various subsampling of endothelial cells from the mouse gastrulation data.

3.4.5. SEACells enables inference of epigenetic regulation

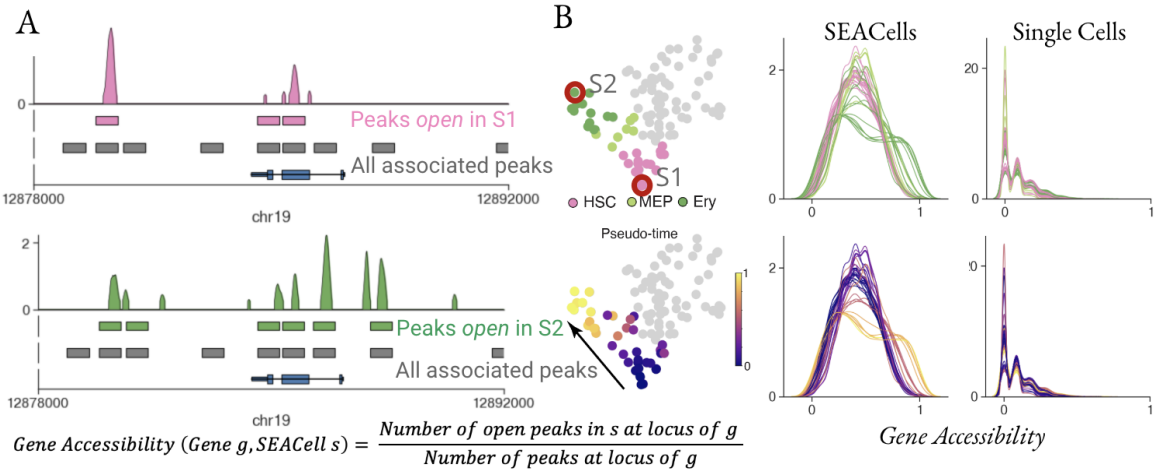
Peaks of ATAC-seq read counts represent open chromatin regions, and gene regulation can be inferred by identifying putative transcription factor (TF) binding motifs within these accessible regions. However, the sparsity of scATAC-seq data has severely restricted its utility, as analysis typically occurs at the resolution of clusters. A typical SEACells metacell contains 1.2 million reads, a large improvement over the 25,000 reads in an individual cell. In order to perform gene regulatory inference, we associate each gene with the specific elements that regulate it. Using SEACell metacells from the ATAC modality of the $CD34^{+}$ bone marrow dataset, we computed correlations between gene expression and peak accessibility for each peak within +/-100 kb of each gene in a core hematopoietic gene set [58]. For all core genes, accessibility of the most correlated peak using ATAC metacells faithfully tracks with gene expression, representing a substantial improvement over correlations identified from the same data at the single-cell level. The inability to detect such correlations from single-cell data means that SEACells is crucial for detecting potential regulatory regions (Figure 3.5).



3.4.5.1. Figure 3.5. A. Spearman correlation between ATAC metacell-aggregated (top) or single-cell (bottom) gene expression and accessibility of the most correlated peak in CD34+ marker genes. B. Accessibility landscape of erythroid factor in hematopoietic stem cells , myeloid-erythroid progenitors and erythroid cells using SEACells (top) or single cells (bottom)

3.4.6. SEACells reveals dynamics of gene accessibility during differentiation

We demonstrate the potential of SEACells metacells for advanced scATAC-seq analysis by examining how the primed and permissive epigenomic landscape of hematopoietic stem cells reconfigures to a landscape with reduced plasticity and developmental potential in differentiated cells. Tracking accessibility dynamics requires overcoming sparsity to identify which regulatory elements are open and accessible. We identified open elements in each ATAC-metacell and computed gene accessibility scores for all highly regulated genes. We observe that the earliest cell type, hematopoietic stem cell, follows a unimodal distribution, suggesting an epigenomic landscape that is poised for hematopoietic gene expression.



3.4.6.1. Figure 3.6. A. Accessibility for a erythroid lineage defining gene increases from stem cell SEACell (S1) to terminal SEACell (S2). B. Chromatin accessibility distribution of highly regulated genes in all metacells along the erythroid lineage. The emergence of bimodality is gradual and continuous. Signal is poorly defined when using single-cell pseudotime bins rather than metacells.

We examine the gene accessibility dynamics of highly regulated genes in each metacell along the pseudo-temporal order and observe that epigenomic reconfiguration is itself gradual and continuous. As cells differentiate along a lineage, genes that define the lineage gain accessibility peaks, while genes that define alternative lineages lose peaks, resulting in bimodality of gene expression in differentiated cells of the erythroid lineage. Using single-cell pseudotime bins instead of metacells does not reveal any bimodality or dynamics, demonstrating that the resolution of SEACells metacells is uniquely suited for capturing dynamics.

4. Inferring Copy Number Variation from scRNA-sequencing

4.1. Introduction

Genetic instability has long been recognized as a hallmark of tumour progression [LL], whereby tumours generate high levels of genetic diversity that facilitate their acquisition of further hallmark capabilities. Genetic instability arises in both the form of point mutations - changes in the nucleotide sequence, as well as in the form of chromosomal instability - a defect that involves loss, gain or rearrangement of chromosomes during cell division. Such chromosomal instability represents a significant source of intra-tumour heterogeneity [70, 18], and acts as a major obstacle to anti-cancer therapy, providing fuel for resistive clones which may later lead to recurrence [18]. Therefore, the identification and quantification of heterogeneous clones within a tumour is of vital clinical importance. Beyond the identification of copy number variation, however, the functional effect of chromosomal alterations on transcriptional processes is not yet well understood. Simultaneous profiling of transcriptional state alongside the identification of chromosomal instability is therefore paramount in order to facilitate the understanding of individuals' varied response to treatments and the development of therapeutic approaches that are tailored to an individual's heterogeneous tumour.

Single-cell RNA sequencing (scRNA-seq) provides an indispensable tool to observe and characterise the full extent of intra-tumour heterogeneity at the transcriptional level, but must be tied to simultaneous measurement of copy number alterations in order to gain a deeper understanding of the functional differences between subclones in a heterogeneous tumour. Currently, parallel sequencing of genomes and transcriptions via scDNA-seq and scRNA-seq respectively remains a challenging endeavour. Furthermore, scDNA-sequencing remains significantly more costly than scRNA-sequencing, and the availability of scRNA-sequencing data is far more widespread. Taken together, this motivates the development of approaches for determining copy number alterations directly from scRNA-seq data, in which variation in gene expression levels can be used to infer variations in copy number between cells.

4.2. Limitations of Previous Approaches

Copy number inference from single-cell data is complicated by the high levels of noise in scRNAseq data and biological and technical dropout which results in sparsity. In order to tackle these issues, existing methods for inference of CNVs from single-cell RNA sequencing such as inferCNV or HoneyBADGER partially rely on averaging the gene expression profiles over large gene segments or over large groups of cells and subsequent detection of large-scale or chromosome arm level copy number aberrations. InferCNV computes a rolling window average expression, which smooths over breakpoints and, consequently, can inaccurately identify them or miss small copy number alterations altogether which affects downstream subclonal inference. HoneyBADGER performs inference using a three-state Hidden Markov Model, modelling gene expression response to copy number as an additive relationship. Existing methods fail to accurately model the relationship between the transcriptional state and copy number profiles of different cell types that are present in the data. The presence of distinct cell types in a tumour will also be reflected in variations in gene expression levels, genetic variations notwithstanding, and neglecting to consider the confounding effect of cell type when inferring copy number from gene expression levels can lead to incorrect inference. Finally, the emergence of intra-tumour heterogeneity arises as a consequence of evolution, which imposes a phylogenetic relationship between subclones. Making use of such phylogenetic relationships whilst accounting for the possibility of independently recurring copy number alterations can allow us to more accurately characterise the sequence of copy number state across genes, or *copy number profiles*, within subclones, and vice versa.

To address these limitations, we developed PICASSO - **P**hylogenetic **I**nference of **C**opy number **A**lterations from **s**c-RNA **S**equencing **O**bservations. PICASSO iteratively infers single-cell copy number profiles as well as phylogenetic relationships between subclones, allowing for cell-type specific gene responses to copy number in order to account for differences in transcriptional programs across cell types.

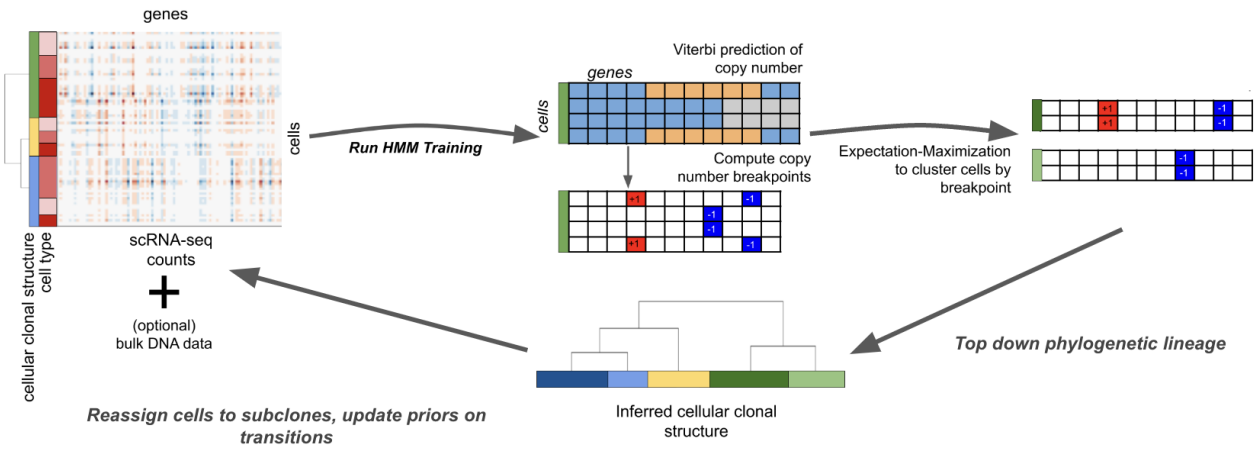
4.3. Method

PICASSO is an algorithm to identify copy number alterations, regions of the genome that exhibit changes to chromosome structure characterised by gains or losses in the number of copies of DNA. The PICASSO algorithm combines the identification of copy number alterations (CNAs) with the construction of a phylogenetic tree in which inferred subclones exhibit different patterns of CNAs. PICASSO assumes that patterns of gene expression observed in a single-cell RNA-sequencing experiment are driven by both gene expression programs characteristic of its particular cell type as well as the underlying copy number profile of that gene. Thus, the PICASSO algorithm:

(1) Aims to infer per-gene copy number profiles from scRNA-seq data, accounting for the combined effects of the relationship between gene expression and the cell's copy number profile and transcriptional program/cell type.
(2) Builds a phylogenetic tree based on inferred patterns of accumulated copy number alterations, and further takes advantage of the priors induced by this evolutionary tree to iteratively refine copy number profile prediction within subclones.

PICASSO comprises 5 main steps:

(1) Initialize parameters of the Hidden Markov Model. The parameters of this HMM include subclone-specific transition matrices that reflect any existing priors on evolutionary structure in the dataset and cell-type specific emission distributions.
(2) Perform Baum-Welch learning to update transition probabilities and parameters of the emission distribution.
(3) Infer copy number profiles for each cell using the Viterbi algorithm.
(4) Split clones into subclones based on probabilistic profiles of copy number breakpoint profiles, using an expectation maximisation algorithm.
(5) Reassign cells to subclones based on likelihood of generating its expression profile from each subclones' copy transition matrix.



4.3.0.1. Figure 4.1. PICASSO uses an iterative algorithm to learn (1) a Hidden Markov Model for inferring copy number alterations and (2) phylogenetic substructure to improve copy number calls.

This pipeline is repeated until the algorithm determines that there is insufficient evidence to further subdivide any clones. PICASSO outputs a copy number profile for each cell, groupings of cells that represent subclones with common copy number profiles and a phylogenetic tree describing the evolutionary relationships between subclones (Figure 4.1).

These design principles allow PICASSO to:

(1) Consider cell-type specific responses of expression to copy number
(2) Increase subclone-specific copy number profile accuracy by constructing priors on each subclone
(3) Examine the evolutionary relationship between subclones from the inferred phylogenetic tree, while taking into account the potential for independently recurring copy number alterations by splitting clones on probabilistic copy number profiles, allowing for the possibility of multiple clonal lineages to contain the same

The major inputs to the PICASSO algorithm are: (i) raw count matrices of gene expression for (potentially) multiple samples, (ii) (optional) cell-type annotations for each cell and (iii) (optional) diploid reference raw count matrices specifying the expected gene expression in diploid cells under each cell type.

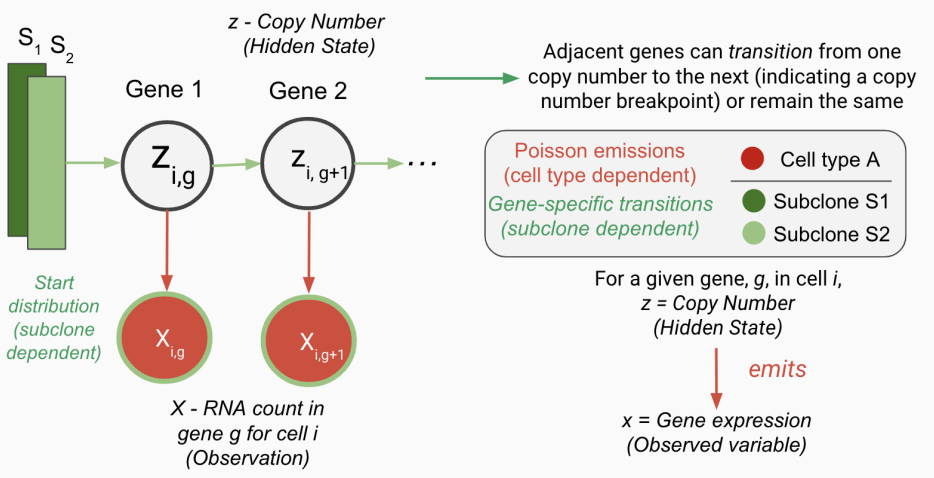
When no cell-type annotations are available, all cells are assigned to the same default cell-type. When no reference is available, a referenceless model can be used where the diploid reference is substituted for the mean gene expression within each cell type. We then infer relative copy number calls which inform the resulting phylogeny.

We denote by $X \in \mathbb{R}^{N \times G}$ the matrix of observed scRNA-sequencing counts, where each row corresponds to a cell and each column corresponds to a gene. Genes are ordered sequentially by chromosomal position. Each cell is annotated with a cell-type and subclone identity. Initially, all cells may be assigned to a single subclone, or any clonal information known *a priori* may be encoded in the model.

4.3.1. Hidden Markov Model for Copy Number Variants

A Hidden Markov Model (HMM) is used to model the single-cell RNA-sequencing counts as a function of the underlying copy number. For each cell, the HMM models an underlying and observed vector of (relative) integer copy number states for each gene, ranging between 0 and K, where K is a user-specified parameter that reflects the range of copy numbers present in the data. By default, K is set to 5, but can be adjusted based on the degree of amplification present in the user's dataset.

Each copy number state is associated with an underlying emission distribution, which specifies the distribution of observed scRNA-seq counts conditioned on the latent integer copy number state. Copy number breakpoints, or changes in inferred copy number state between adjacent chromosomal regions, are identified through inference on the matrix of transition probabilities between successive genes. This probabilistic model is illustrated in Figure 4.2.



4.3.1.1. Figure 4.2. Cell-type and subclone specific Hidden Markov Model used to infer latent copy number state from observed single cell RNA-sequencing counts.

Cell-type Specific Emission Distribution

A key innovation of PICASSO is the observation that multiple factors beyond copy number influence the level of expression of a particular gene in a given cell. Inference of copy number from gene expression is

predicated on the assumption that variation in scRNA-seq counts reflect differences in the number of copies of the gene, but should not neglect that additional factors such as cell-type specific transcriptional programs and cell-dependent library size may obscure this signal. In order to account for this, PICASSO models the emission of counts given a copy number state as a Poisson distribution, with a cell-type specific mean which is also modified by cell-dependent library size.

Formally, for cell i belonging to cell type C, the distribution of counts, $x_{i,g}$, for gene g, conditioned on the latent copy number state, $z_{i,g}$ of that (cell, gene) pair follows a Poisson distribution as follows:

$$P(x_{i,g}|z_{i,g} = k) \sim Poisson\,(m_i \times \lambda^{(2)}_{g,C} \times [\tfrac{k}{2}]^{\alpha_{g,C}}) \tag{4.1}$$

The first term, m_i - computed as the ratio of the total library size of that cell to the library size of the average diploid cell of that cell type - is a cell-specific normalisation term to control for the influence of biological and technical variation between cells on predicted copy number.

Gene Copy Number Response

Gene copy number response, $[\frac{k}{2}]^{\alpha_{g,C}}$, models a power law relationship between the expected observed count and underlying copy number state, where the exponent, $\alpha_{g,C}$, is a learned gene- and cell-type specific parameter that facilitates the modelling of non-linear relationships between copy number and expression. Gene copy number response is optimised via Baum-Welch, with a prior of $\mathcal{N}(1, \sigma^2)$ to reflect the expectation of near-linear response.

Diploid Reference

The diploid reference expression of gene g in cell type C, $\lambda^{(2)}_{g,C}$, specifies the expected gene expression in the diploid state for each cell type. It is used to compare observed expression in a (potentially) copy number altered cell to a baseline diploid expression level. In the absence of reference data for each cell-type, the same diploid reference may be used for multiple cell types.

Transition Probability Matrix

PICASSO assumes that copy number breakpoints are relatively infrequent events, enforced through the structure of the transition matrix as follows. The transition probabilities between states j and k at the junction between gene g and gene $g + 1$ are defined as follows:

$$T(j,k) = \begin{cases} 1 - \epsilon_g & if\ k = j + \Delta_g \\ \epsilon_g & otherwise \end{cases} \tag{4.2}$$

where ϵ_g denotes a small probability representing unlikely transitions. This is initialised at a small value $\epsilon_g = 0.01$ and is fit to observed transitions during optimization. Δ_g is an offset representing the

magnitude of the most likely transition. *A priori*, this offset is equal to zero, reflecting the assumption that copy number alterations are relatively rare. Following a round of learning where copy numbers are inferred by the Hidden Markov Model, this offset is updated to reflect the learned copy number change in a given subclone. For example, if a breakpoint is inferred at gene g, with a copy number change from diploid (2) to triploid (3), the offset is updated to $\Delta_g = 1$.

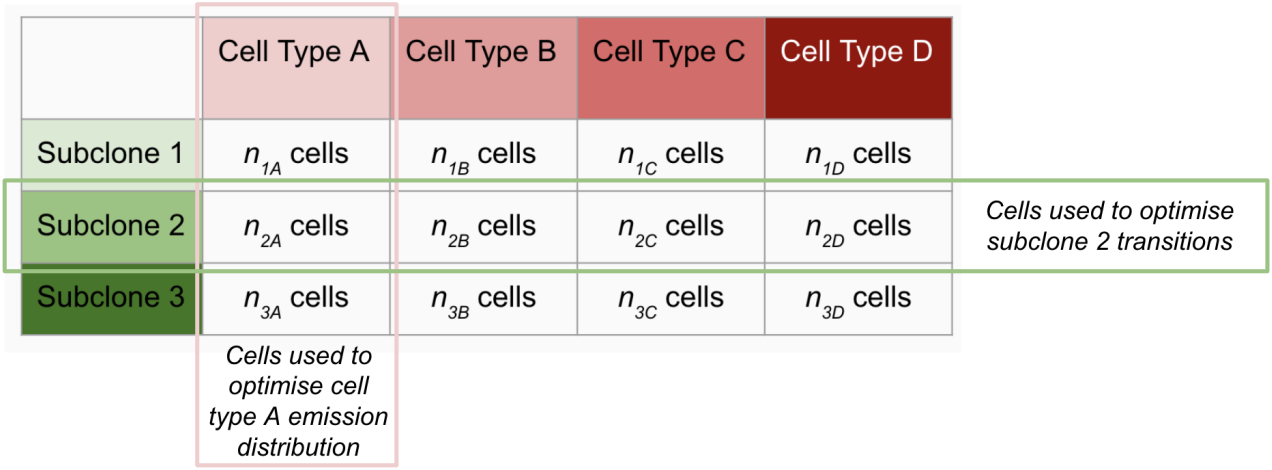
Hidden Markov Model Optimisation

For each (subclone, cell-type) pair, we construct a Hidden Markov Model as illustrated in Figure 4.2. The Baum-Welch algorithm is used to infer (1) gene-specific copy number response, $\alpha_{g,C}$ and (2) subclone-specific transition probabilities $\epsilon = \{\epsilon_1, \cdots \epsilon_G\}$. However, in order to improve the power of the model:

(1) Gene-specific copy number response (which does not depend on subclone identity) is updated using all cells in a given cell type, potentially across multiple subclones.

(2) Subclone-specific transition probabilities (which do not depend on cell-type specific emission distributions) are updated using all cells of a given subclone, potentially across multiple cell-types.

This relationship is highlighted in Figure 4.3 below.



4.3.1.2. Figure 4.3. Pooling cells across different cell types and different subclones in Hidden Markov Model enables increased inference power from larger effective data size.

4.3.2. Phylogeny and Subclone Identification

We leverage our inference of copy number calls from scRNA sequencing data to reconstruct a tumour phylogeny to gain insight about the high level relationships between clones as well as use the phylogeny to further refine copy number predictions. The phylogeny is initialised with a single clone containing all the cells in the data set. At each iteration, the depth of the existing tree may be increased by one by splitting a leaf clone into two further subclones. The Expectation Maximisation algorithm is used to cluster each clone into two subclones using an expectation maximisation algorithm by breakpoint profiles. The

breakpoint profile is defined as the sign of the difference in copy number between successive genes, or 0 if there is no change between genes.
This approach allows us to capture two important features by learning categorical probabilities for each subclone.

(1) The learned probability associated with the categorical distribution for a given breakpoint can be less than 1, permitting an event in which a breakpoint is present only in a subset of cells in an inferred subclone which can be further disentangled by subclone splitting in the subsequent iteration.
(2) There may be a positive probability of a particular breakpoint occurring at the same position in both clones, allowing for the independent recurrence of copy number changes in multiple clones.

4.4. Results

4.4.1. Metrics

We have evaluated PICASSO using the following metrics:

(1) AUC Score (Area Under the ROC Curve) - measures a classifier's power to separate classes.
(2) Accuracy - proportion of instances where the inferred labels match the ground truth label.

4.4.2. Datasets

In the absence of a ground truth dataset, we evaluate our approach on a combination of (1) aligned single-cell DNA sequencing/ single-cell RNA sequencing of two clonally related high grade serous carcinoma cell lines[14] (2) simulations using Splatter [78], adapted to include expression change with copy number alterations and (3) *in silico* copy number altered chromosomes, constructed by stitching diploid regions and regions of known copy number alterations (from bulk data from the same sample) to construct clones with copy number alterations of varying widths and to construct phylogenies with mixed tumour and normal clones.

4.4.3. Incorporating cell type specific parameters improves inference

We first investigate whether incorporating cell-type specific information improves accuracy of copy number calls, using Splatter simulations. We compare both the clonal accuracy and copy number call accuracy of PICASSO with cell type information and without it, and find that incorporating this information significantly improves the accuracy of our model, suggesting that modelling these features is important for accurate inference of copy number from RNA expression data.

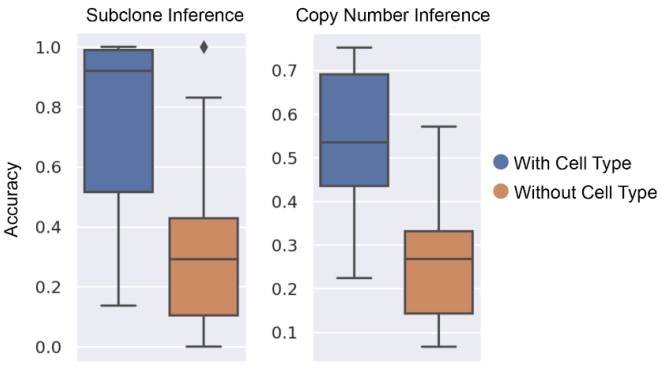


Figure 4.4. Incorporating cell type information improves accuracy of subclone and copy number inference

4.4.4. Incorporating subclone-specific priors improves CNV accuracy

Next, we examine whether incorporating subclone-specific priors improves accuracy of copy number prediction by evaluating copy number accuracy over PICASSO iterations. We find that over three iterations of learning the phylogenetic structure, PICASSO copy number prediction accuracy improves, indicating that subclone splitting using Expectation Maximisation and incorporating subclone-specific priors improves accuracy of copy number prediction. PICASSO was able to recover subclones at frequencies up to 10% of the total dataset, and generally outperformed HoneyBADGER which was often unable to distinguish tumour from non-tumour cells.

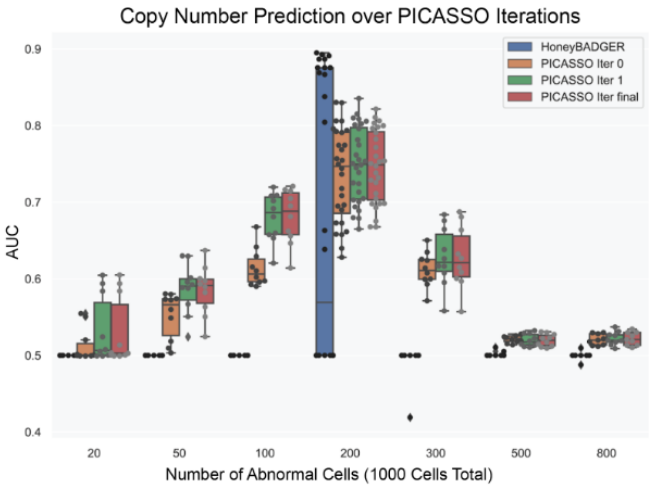


Figure 4.5. Incorporating phylogenetic structure improves PICASSO predictions each iteration

4.4.5. PICASSO outperforms competing methods across scales of copy number variation

We quantify PICASSO's ability to detect copy number alterations of varying widths. InferCNV does not provide copy numbers directly, so we performed comparisons against HoneyBADGER only. PICASSO provides two options for copy number - the direct copy number profile of each cell as well as a per-subclone copy number profile which smooths over the noise present in single cells. We constructed a simulated chromosome of 1000 genes and inserted regions of known copy number alteration of varying widths. We find that PICASSO generally outperforms HoneyBADGER, and is particularly able to recover focal copy number alterations. At larger widths, the performance of PICASSO worsens as the assumption of an on-average diploid chromosome no longer holds.

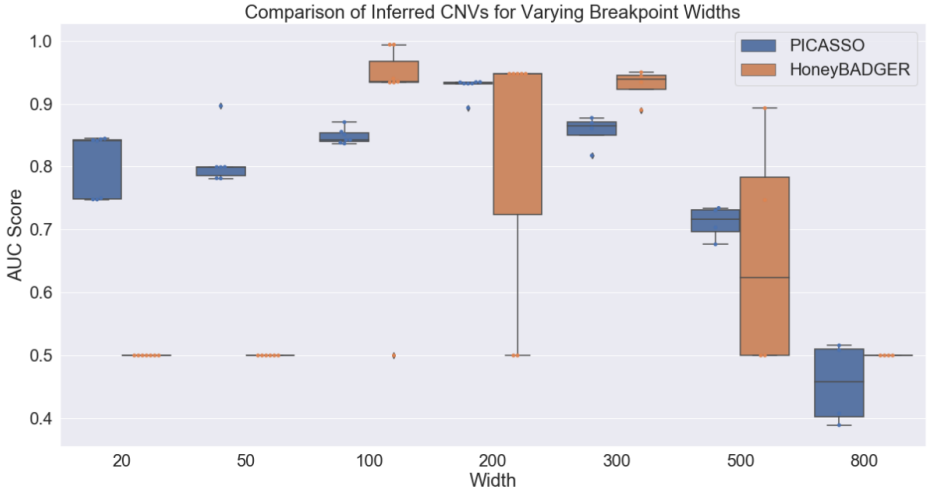


Figure 4.6. AUC score for inferred copy number profiles for regions of different copy number altered width

4.4.6. PICASSO outperforms competing methods at distinguishing tumour cells

In order to assess PICASSO on true biological data, we used scRNA-sequencing data aligned to distinct clones derived from scDNA-sequencing in the same sample. We compare the clonal accuracy of each method and find that HoneyBADGER is significantly worse at recovering clonal identity. PICASSO outperforms all methods, even when employing a referenceless model which computes relative copy numbers (Figure 4.7).

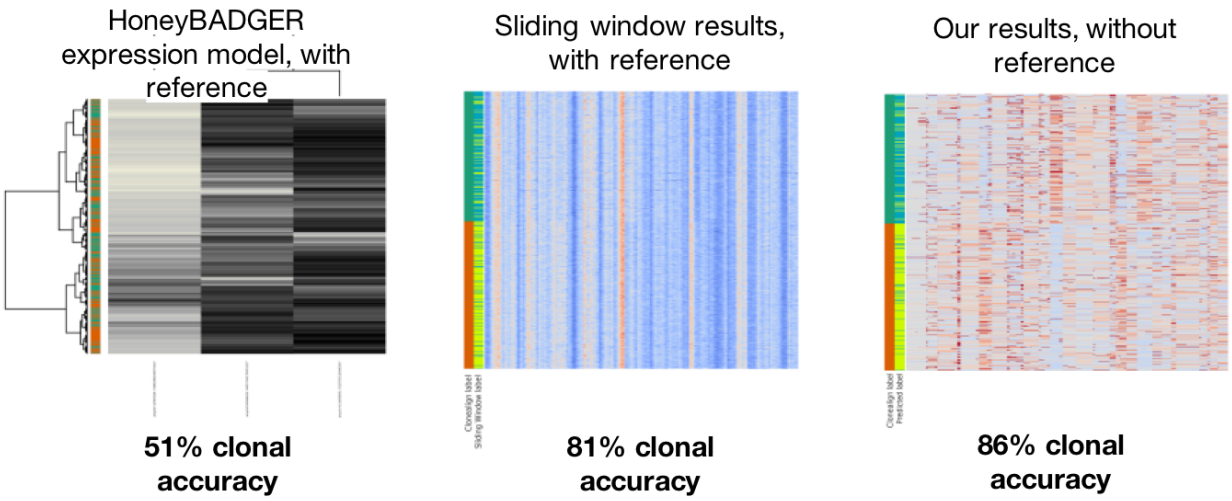


Figure 4.7. Accuracy of inferred clones (tumour/non-tumour) from high grade serous carcinoma cell lines across methods.

4.5. Proposed Work

Reconstruction of copy number alterations from single-cell data is an inherently difficult task due to noise in the dataset. In order to mitigate this, we propose first applying SEACells (Section 3) to the input data, and passing these aggregated metacells as input to PICASSO. This will have the dual effects of (1) reducing noise and (2) reducing the scale of the data. We anticipate that differences in copy numbers will be reflected in gene expression levels, and thus different subclones will aggregate in different SEACells. Therefore we propose the following future work:

(1) Simulate datasets with multiple subclones and cell-types using Splatter. We will apply SEACells to these datasets in order to verify that SEACells are of high cell-type purity and high clonal purity
(2) Run PICASSO with SEACells inputs and benchmark the performance of this combined method using the existing datasets we describe in Section 4.4.

Additional work to complete this project includes:

(1) Complete benchmarking on more recently published copy number inference methods, such as CasPER[57] and CopyKAT [22].
(2) Incorporate balanced accuracy scores to measure performance on imbalanced datasets.

5. Proposed Work - Hyperbolic Embeddings for Reconstruction and Analysis of Lineages

5.1. Hyperbolic Embeddings for CRISPR-Cas9 Lineage Tracing

5.1.1. CRISPR-Cas9 Evolutionary Model

CRISPR-Cas9 lineage tracing experiments involve the insertion of black 'scratchpads' consisting of multiple blank target sites into a model system. Heritable CRISPR-induced mutations accumulate at target sites over generations, with the ultimate goal of sequencing patterns of mutations in offspring that can be used to reconstruct their evolutionary history, using single-cell RNA sequencing.

Canonical phylogenetic algorithms which construct distance matrices for traditional mutation data may not be suitable for the analysis of lineage tracing data, which are generated from very different model assumptions and very different mechanisms of noise. Therefore, the construction of a distance matrix from lineage tracing data must take into account specific features of the lineage tracing system.

Traditional mutation data follows a reversible mutation model, where any pair of nucleotide transition is possible, and back mutations are allowed. A commonly used evolutionary models, the Jukes-Cantor model is illustrated in Figure 5.1:

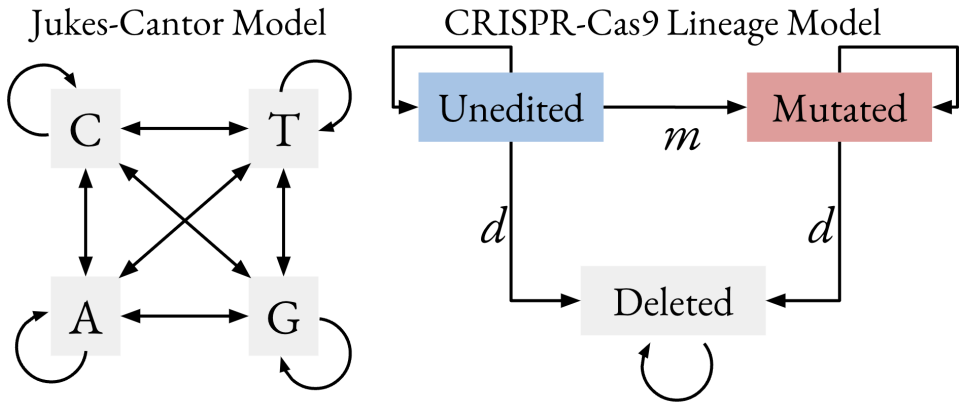


Figure 5.1. Traditional models of evolution (left) follow a very different transition matrix from CRISPR lineage systems

The evolutionary model for a CRISPR lineage tracing is quite different from these canonical models; to start, the initial state of all target states is known to be the unedited state, by cassette design. Once a mutation occurs at a target site, no further edits can be made, except in the case where a neighbouring site induces a deletion. There are multiple potential mutated states, which we have condensed in the above illustration for simplicity, as transitions between mutated states are not possible in a CRISPR-Cas9 system.

Some key differences between this model and canonical evolutionary models are:

1. Time-irreversibility - once an edit is made, reversions to an unedited start or other mutations are not possible.
2. Site-specific mutation rates - in order to capture evolutionary relationships at varying time-scales, recording systems are often designed with multiple guides that induce CRISPR-Cas9 edits at widely varying rates, which must be taken into account when computing distances between cells
3. Known ancestral state - unlike DNA sequences, the ancestral state, or root of the lineage tree is known to be the unedited state. This information can be used to inform the construction of the phylogenetic tree.

Beyond differences in the evolutionary model, the single-cell RNA sequencing data generated from lineage tracing systems have different characteristics from mutation data inferred from bulk DNA sequencing, for example.

1. Noisy sequencing of indels - mutation states correspond to unique indels generated by CRISPR cuts in the DNA, not single nucleotide edits. Noisy sequencing of bases may cause identical mutation states to be incorrectly identified as two distinct mutations, unless information about similarities between indels are used to inform mutation calls.
2. Recurrent mutations and non-random indel insertion - CRISPR inserts random indels into target sites. However, these indels do not occur with uniform probability, and certain indels may be preferentially inserted. This may lead to mutations recurring independently across different branches of the evolutionary tree.
3. Missing data - if target sites are not sequenced for a given cell in the sequencing experiment, this data is missing, and must be incorporated into the reconstruction algorithm.
4. Large scale deletions - CRISPR edits may cause large scale indels which modify or delete neighbouring target sites, thus affecting their mutation profiles.

Altogether, these features motivate the development and optimization of an evolutionary model specific to a CRISPR-Cas9 lineage tracing in order to more accurately reconstruct a phylogenetic tree. We propose to learn distances between cells by optimising the log-likelihood of our observed data under a customised evolutionary model.

5.1.2. Optimization in Hyperbolic Space

As described in Section 2.4.3, hyperbolic distance metrics possess certain properties which make them well suited for tree construction. Namely, distances in hyperbolic space satisfy a relaxed version of the four point condition, which characterises interleaf distances on trees. This motivates the optimization of cell-cell distances in hyperbolic space to more faithfully reconstruct a lineage tree. We propose to perform this optimization using Riemannian gradient descent in the hyperboloid model of hyperbolic space, which has convenient formulae for computing the gradient of the distance function [72].

5.2. Proposed Approach

HERACLES - **H**yperbolic **E**mbeddings for **R**econstruction and **A**nalysis of **C**RISPR-Cas9 **L**ineag**ES** - is our proposed algorithm for inferring lineage tree distances between sequenced cells. The HERACLES algorithm optimises distances between cells in hyperbolic space, using Riemannian stochastic gradient descent, in order to maximise a custom CRISPR lineage tracing experiment-specific likelihood function. HERACLES defines a likelihood function to model transition probabilities between mutation states at each target site that takes into account (1) different mutation rates at each sites, (2) non-uniform distributions over indels at each target site and (3) the time-irreversible evolutionary model of lineage tracing experiments. The full derivation of the likelihood function of the data is described in Appendix C.

The major inputs to the HERACLES algorithm are:
(i) A character matrix which defines the mutation state of each target site in each cell. The mutation state can either be a single mutation character, or a distribution over possible mutation states to incorporate uncertainty in sequencing data. In the case of missing or deleted data, we may impute likely mutation states from similar cells.
(ii) (Optional) Site-specific mutation and deletion rates. If not provided, it can be learned from the data.
(iii) (Optional) CRISPR indel distributions. If not provided, it can be learned from the data.

We first embed all cells in hyperbolic space in an initial point configuration, then iteratively update the point configuration of all cells in order to maximise our data likelihood. The output of this iterative process is a distance matrix which can be used to probe relationships between cells, and thus construct a lineage tree or address questions of relatedness between different cells and their respective patterns of transcription.

The proposed algorithm comprises 5 main steps:

(1) Construct an initial 'best guess' tree based on ancestry-aware Hamming distance between cells.
(2) Use Felsenstein's pruning algorithm [20] to compute likely internal states, and, if not specified by the user, estimate mutation and deletion rates for each target site, as well as estimate probability distributions over indels at each site.
(3) Embed this tree in hyperbolic space [50] to obtain an initial point configuration.
(4) Iteratively update each cell's position in hyperbolic spaces using Riemannian gradient descent in order to maximise the likelihood of the data.
(5) Use distance matrix from the learned point configuration to reconstruct an updated lineage tree.

This pipeline is repeated until the likelihood of the data converges.

5.3. Preliminary Results

In order to assess the feasibility of this method for reconstructing lineage trees, we have performed comparisons between tree distances which were inferred from the customised likelihood function optimised in Euclidean space. We compare the Robinson-Foulds distance between our ground truth simulated tree and our inferred tree, as well as the distances between existing methods' trees and the inferred tree. (Figure 5.2)

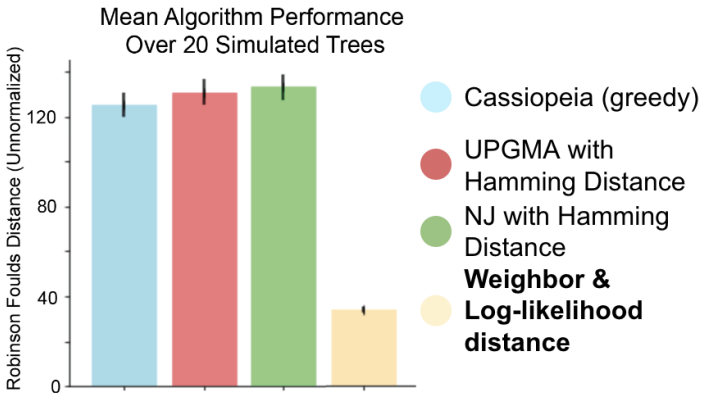


Figure 5.2 - Log-likelihood distance with weighbor tree reconstruction improves fidelity of inferred tree to simulated tree

5.4. Benchmarking and Evaluation

5.4.1. Metrics

We plan to benchmark this approach using a combination of real and simulated data. We will primarily assess tree reconstruction using the following metrics:

(1) Triplets distance - Every triplet of leaves is samples and their inferred lineage relationship is compared to the ground truth lineage relationship.

(2) Robinson-Foulds distance [11]- Every possible partition of leaves is computed and the number of shared partitions between the inferred tree and ground truth tree are counted.

Both metrics are normalised to obtain a value between 0 (perfect agreement) or 1 (comparable to a random guess). For real experimental data where no ground truth tree is known, we rely on biological evaluation to compare algorithms.

5.4.2. Datasets

CRISPR-Cas9 lineage tracing is a recent experimental development, and as such, there are not many high-quality biological datasets readily available, although the number is rapidly growing. One recent paper [75] published a months-long continuous cell lineage tracing of lung adenocarcinoma in mice and contains lineage tracing data for 40,386 cells detected across 35 tumours. The DREAM challenge for lineage reconstruction [24] has also made available *in silico* data for a C. elegans lineage tree of approximately 1,000 cells and a Mus musculus tree of 10,000 cells. Beyond these publicly available datasets, we have already written simulation code for generating *in silico* tumour lineage tracing data, which we will use for further algorithmic development and benchmarking.

6. Research Plan

The plan for the completion of this dissertation is as described in Table 6.1 below.

Timeline	Work	Progress
Fall 2018 - Fall 2021	Design and implementation of PICASSO - **P**hylogenetic **I**nference of **C**opy number **A**lterations from **s**c-RNA **S**equencing **O**bservations	Complete
Spring 2021 - Fall 2022	SEACells - Inference of transcriptional and epigenomic cellular states from single-cell genomics data *Under Review at Nature Biotechnology*	Complete
Fall 2022 - Spring 2023	Completion of comprehensive benchmarking of PICASSO on single-cell data	In progress
Fall 2022 - Spring 2023	Apply SEACells aggregation approach to PICASSO for more robust inference of copy number alterations	In progress
Fall 2022 - Summer 2023	Implementation and benchmarking of hyperbolic embeddings for CRISPR-Cas9 lineage tracing	In progress
Spring 2023 - Fall 2023	Thesis Writing	To do
Fall 2023	Thesis Defence	To do

Table 6.1. Timeline for competition of proposed and ongoing work.

7. References

1. Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., ... & Zucman-Rossi, J. (2015). Toward understanding and exploiting tumor heterogeneity. Nature medicine, 21(8), 846-853. https://www.sciencedirect.com/science/article/pii/S153561082100386X
2. Altschuler, S. J., & Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference?. Cell, 141(4), 559–563. https://doi.org/10.1016/j.cell.2010.04.033
3. Andrews TS and Hemberg M. False signals induced by single-cell imputation [version 2; peer review: 4 approved]. F1000Research 2019, 7:1740 (https://doi.org/10.12688/f1000research.16613.2)
4. Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., & Theis, F. J. (2017). Single cells make big data: New challenges and opportunities in transcriptomics. Current Opinion in Systems Biology, 4, 85-91.
5. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., & Garmire, L. X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome biology, 20(1), 1-14.
6. Azizi, E., Prabhakaran, S., Carr, A., & Pe'er, D. (2017). Bayesian inference for single-cell clustering and imputing. Genomics Comput Biol, 3(1), 46.
7. Baran, Y. et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol 20, 206 (2019). https://doi.org:10.1186/s13059-019-1812-2
8. Bauckage, C., Kersting, K., Hoppe, F. & Thurau, C. in Workshop New Challenges in Neural Computation (2015).
9. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol 23, 100 (2022). https://doi.org:10.1186/s13059-022-02667-1
10. Bilous, M. et al. Super-cells untangle large and complex single-cell transcriptome networks. bioRxiv, 2021.2006.2007.447430 (2021). https://doi.org:10.1101/2021.06.07.447430
11. Briand, S., Dessimoz, C., El-Mabrouk, N., Lafond, M., & Lobinska, G. (2020). A generalized Robinson-Foulds distance for labeled trees. BMC genomics, 21(10), 1-13.
12. Buenrostro, J., Wu, B., Litzenburger, U. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015). https://doi.org/10.1038/nature14590
13. Camin, J. H., & Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. Evolution, 311-326.
14. Campbell, K. R., Steif, A., Laks, E., Zahn, H., Lai, D., McPherson, A., ... & Shah, S. P. (2019). clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome biology, 20(1), 1-12.
15. Carter, B., Zhao, K. The epigenetic basis of cellular heterogeneity. Nat Rev Genet 22, 235–250 (2021). https://doi.org/10.1038/s41576-020-00300-0
16. Chen, M., & Zhou, X. (2018). VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome biology, 19(1), 1-15.
17. Cutler, A. & Breiman, L. Archetypal analysis. Technometrics 36.4, 338-347 (1994).
18. Dagogo-Jack, I., Shaw, A. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15, 81–94 (2018). https://doi.org/10.1038/nrclinonc.2017.166
19. Fan, J., Lee, H. O., Lee, S., Ryu, D. E., Lee, S., Xue, C., ... & Kharchenko, P. V. (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome research, 28(8), 1217-1227.
20. Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution, 1229-1242.
21. Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer, 108(3), 479–485. https://doi.org/10.1038/bjc.2012.581
22. Gao, R., Bai, S., Henderson, Y.C. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 39, 599–608 (2021). https://doi.org/10.1038/s41587-020-00795-2
23. Genomics, X. PBMC CITE-seq from a Healthy Donor, <https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0>

24. Gong, W., Granados, A. A., Hu, J., Jones, M. G., Raz, O., Salvador-Martínez, I., ... & Meyer, P. (2021). Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of C. elegans and M. musculus developmental trees. Cell systems, 12(8), 810-826.
25. Gong, W., Kim, H. J., Garry, D. J., & Kwak, I. Y. (2022). Single cell lineage reconstruction using distance-based algorithms and the R package, DCLEAR. BMC bioinformatics, 23(1), 1-14.
26. Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., & Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. BMC bioinformatics, 19(1), 1-10.
27. Gough, A., Stern, A. M., Maier, J., Lezon, T., Shun, T. Y., Chennubhotla, C., Schurdak, M. E., Haney, S. A., & Taylor, D. L. (2017). Biologically Relevant Heterogeneity: Metrics and Practical Insights. SLAS discovery : advancing life sciences R & D, 22(3), 213–237. https://doi.org/10.1177/2472555216682725
28. Guo, M., Peng, Y., Gao, A. et al. Epigenetic heterogeneity in cancer. Biomark Res 7, 23 (2019). https://doi.org/10.1186/s40364-019-0174-y
29. Gupta, R. K., & Kuznicki, J. (2020). Biological and Medical Importance of Cellular Heterogeneity Deciphered by Single-Cell RNA Sequencing. Cells, 9(8), 1751. https://doi.org/10.3390/cells9081751
30. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989-2998 (2015). https://doi.org:10.1093/bioinformatics/btv325
31. Hanahan, D. (2022). Hallmarks of cancer: new dimensions. Cancer discovery, 12(1), 31-46. https://aacrjournals.org/cancerdiscovery/article/12/1/31/675608/Hallmarks-of-Cancer-New-DimensionsHallmarks-of
32. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. cell, 144(5), 646-674.
33. Haque, A., Engel, J., Teichmann, S.A. et al. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 9, 75 (2017). https://doi.org/10.1186/s13073-017-0467-4
34. Hie, B., Cho, H., DeMeo, B., Bryson, B., & Berger, B. (2019). Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell systems, 8(6), 483-493.
35. Ji Z., Zhou W., Ji H.Single-cell regulome data analysis by SCRAT. Bioinformatics. 2017; 33(18):2930–2.
36. Ji, Z., Zhou, W., Hou, W. et al. Single-cell ATAC-seq signal extraction and enhancement with SCATE. Genome Biol 21, 161 (2020). https://doi.org/10.1186/s13059-020-02075-3
37. Jones, M.G., Khodaverdian, A., Quinn, J.J. et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. Genome Biol 21, 92 (2020). https://doi.org/10.1186/s13059-020-02000-8
38. Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020). https://doi.org/10.1186/s13059-020-1926-6
39. Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature communications, 9(1), 1-9.
40. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature methods, 15(12), 1053-1058.
41. Matsumoto, H., Mimori, T., & Fukunaga, T. (2021). Novel metric for hyperbolic phylogenetic tree embeddings. Biology Methods and Protocols, 6(1), bpab006.
42. McKenna, A., & Gagnon, J. A. (2019). Recording development with single cell dynamic lineage tracing. Development, 146(12), dev169730.
43. Method of the Year 2013. Nat Methods 11, 1 (2014). https://doi.org/10.1038/nmeth.2801
44. Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems, 30.
45. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., ... & Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344(6190), 1396-1401.
46. Perkel, J. M. (2021). Single-cell analysis enters the multiomics age. Nature, 595(7868), 614-616. https://www.nature.com/articles/d41586-021-01994-w
47. Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome biology, 16(1), 1-10.
48. Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., ... & Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature, 566(7745), 490-495.

49. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al.Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018; 71(5):858–71.
50. Rik Sarkar. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In Proceedings of the 19th International Conference on Graph Drawing, GD'11, pages 355–366, Berlin, Heidelberg, 2012. Springer-Verlag.
51. Ross, E.M., Markowetz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol 17, 69 (2016). https://doi.org/10.1186/s13059-016-0929-9
52. Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., ... & Hanlon, S. (2020). The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell, 181(2), 236-249.
53. Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., & Teichmann, S. A. (2017). The Human Cell Atlas: from vision to reality. Nature, 550(7677), 451-453.
54. Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
55. Sánchez-Rivera, F. J., Diaz, B. J., Kastenhuber, E. R., Schmidt, H., Katti, A., Kennedy, M., Tem, V., Ho, Y. J., Leibold, J., Paffenholz, S. V., Barriga, F. M., Chu, K., Goswami, S., Wuest, A. N., Simon, J. M., Tsanov, K. M., Chakravarty, D., Zhang, H., Leslie, C. S., Lowe, S. W., ... Dow, L. E. (2022). Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants. Nature biotechnology, 40(6), 862–873. https://doi.org/10.1038/s41587-021-01172-3
56. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017; 14(10):975.
57. Serin Harmanci, A., Harmanci, A.O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun 11, 89 (2020). https://doi.org/10.1038/s41467-019-13779-x
58. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol 37, 451-460 (2019). https://doi.org:10.1038/s41587-019-0068-4
59. Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., ... & Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. Nature biotechnology, 34(6), 637-645.
60. Spanjaard, B., Hu, B., Mitic, N., & Junker, J. P. (2017). Massively parallel single cell lineage tracing using CRISPR/Cas9 induced genetic scars. bioRxiv, 205971.
61. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nature biotechnology, 36(5), 469–473. https://doi.org/10.1038/nbt.4124
62. Stegle, O., Teichmann, S. & Marioni, J. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 16, 133–145 (2015). https://doi.org/10.1038/nrg3833
63. Tsompana, M., Buck, M.J. Chromatin accessibility: a window into the genome. Epigenetics & Chromatin 7, 33 (2014). https://doi.org/10.1186/1756-8935-7-33
64. Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. American journal of human genetics, 85(2), 142–154. https://doi.org/10.1016/j.ajhg.2009.06.022
65. van Dijk D, Sharma R, Nainys J, et al.: Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174(3):716–729.e27. 10.1016/j.cell.2018.05.061
66. van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell 174, 716-729 e727 (2018). https://doi.org:10.1016/j.cell.2018.05.061
67. van Dijk, E., van den Bosch, T., Lenos, K.J. et al. Chromosomal copy number heterogeneity predicts survival rates across cancers. Nat Commun 12, 3188 (2021). https://doi.org/10.1038/s41467-021-23384-6
68. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. science, 339(6127), 1546-1558.
69. Wagner F, Yan Y, Yanai I: K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. bioRxiv. 2017. 10.1101/217737

70. Walker, B. A., Wardell, C. P., Melchor, L., Brioli, A., Johnson, D. C., Kaiser, M. F., ... & Morgan, G. J. (2014). Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. Leukemia, 28(2), 384-390.
71. Wilson, B. (2021). Learning phylogenetic trees as hyperbolic point configurations. arXiv preprint arXiv:2104.11430.
72. Wilson, B., & Leimeister, M. (2018). Gradient descent in hyperbolic space. arXiv preprint arXiv:1805.08207.
73. Wu, A., Neff, N., Kalisky, T. et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods 11, 41–46 (2014). https://doi.org/10.1038/nmeth.2694
74. Wu, S. H. S., Lee, J. H., & Koo, B. K. (2019). Lineage tracing: computational reconstruction goes beyond the limit of imaging. Molecules and cells, 42(2), 104.
75. Yang, D., Jones, M. G., Naranjo, S., Rideout III, W. M., Min, K. H. J., Ho, R., ... & Weissman, J. S. (2022). Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell, 185(11), 1905-1923.
76. Yao, M., Ren, T., Pan, Y., Xue, X., Li, R., Zhang, L., Li, Y., & Huang, K. (2022). A New Generation of Lineage Tracing Dynamically Records Cell Fate Choices. International journal of molecular sciences, 23(9), 5021. https://doi.org/10.3390/ijms23095021
77. Zafar, H., Tzen, A., Navin, N. et al. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biol 18, 178 (2017). https://doi.org/10.1186/s13059-017-1311-2
78. Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. Genome biology, 18(1), 1-15.
79. Zhao C, Hu S, Huo X, Zhang Y. Dr. seq2: a quality control and analysis pipeline for parallel single cell transcriptome and epigenome data. PLoS One. 2017; 12(7):0180583.

Appendix A: SEACells Kernel Matrix Construction

Nearest Neighbour Graph

A k-nearest neighbour graph is constructed using Euclidean distance in the low-dimensional embedding (PCA or SVD). The graph is composed of nodes representing single cells, each connected to their most similar neighbours. The nearest graph can be represented as a matrix $D \in R^{n\,X\,n}$, where n is the number of cells. D_{ij} represents distance between cells i and j if they are neighbours and $D_{ij} = 0$ otherwise.

Construction of the affinity kernel matrix

The goal of SEACells algorithm is to identify metacells composed of tightly related groups of cells. Therefore, we transform the distances in the neighbour graph to similarities between neighbouring cells. Gaussian kernels provide a typical approach for this transformation, but assume that densities in underlying data are approximately uniform. Single-cell data, however, shows remarkable variability in data densities with low-density regions or rare cell-types often the most meaningful in describing the biology of the system. An adaptive kernel that uses neighbour distance as the scaling factor for each cell, rather than a fixed parameter, is highly effective in faithfully representing the phenotypic similarities [QQQ, RRR]. The adaptive kernel corrects for densities using distance to the l^{th} $(l < k)$ nearest neighbour as a scaling factor, i.e. the scaling factor of cell i is given by σ_i =distance to l^{th} nearest neighbour.

The adaptive Gaussian kernel is then given by

$$M(x_i, x_j) = \frac{1}{\sqrt{2\pi(\sigma_i + \sigma_j)\ )}} \exp\left(-\frac{1}{2}\frac{(x_i - x_j)^T(x_i - x_j)}{\sigma_i + \sigma_j}\right)$$

where x_i is the low-dimensional embedding corresponding to cell i i.e. PCA for single-cell RNA-seq and SVD for single-cell ATAC-seq. $M \in R^{n\,X\,n}$ is the affinity matrix. M_{ij} represents the similarity between cells i and j if they are mutual neighbours and $M_{ij} = 0$ otherwise and n is the number of cells.

In this kernel space, two cells (x and y) are embedded close to each other if they satisfy two conditions:

3. x and y share neighbours in the PCA/SVD space
4. the similarity score among the neighbours of x and y are similar

Two cells in this transformed dimensional space will be similar to each other not only if they share the neighbours but only if the distances to the shared neighbours are also similar, imposing stricter similarity conditions between cells.

Appendix B: SEACells Benchmarking Against Previous Methods

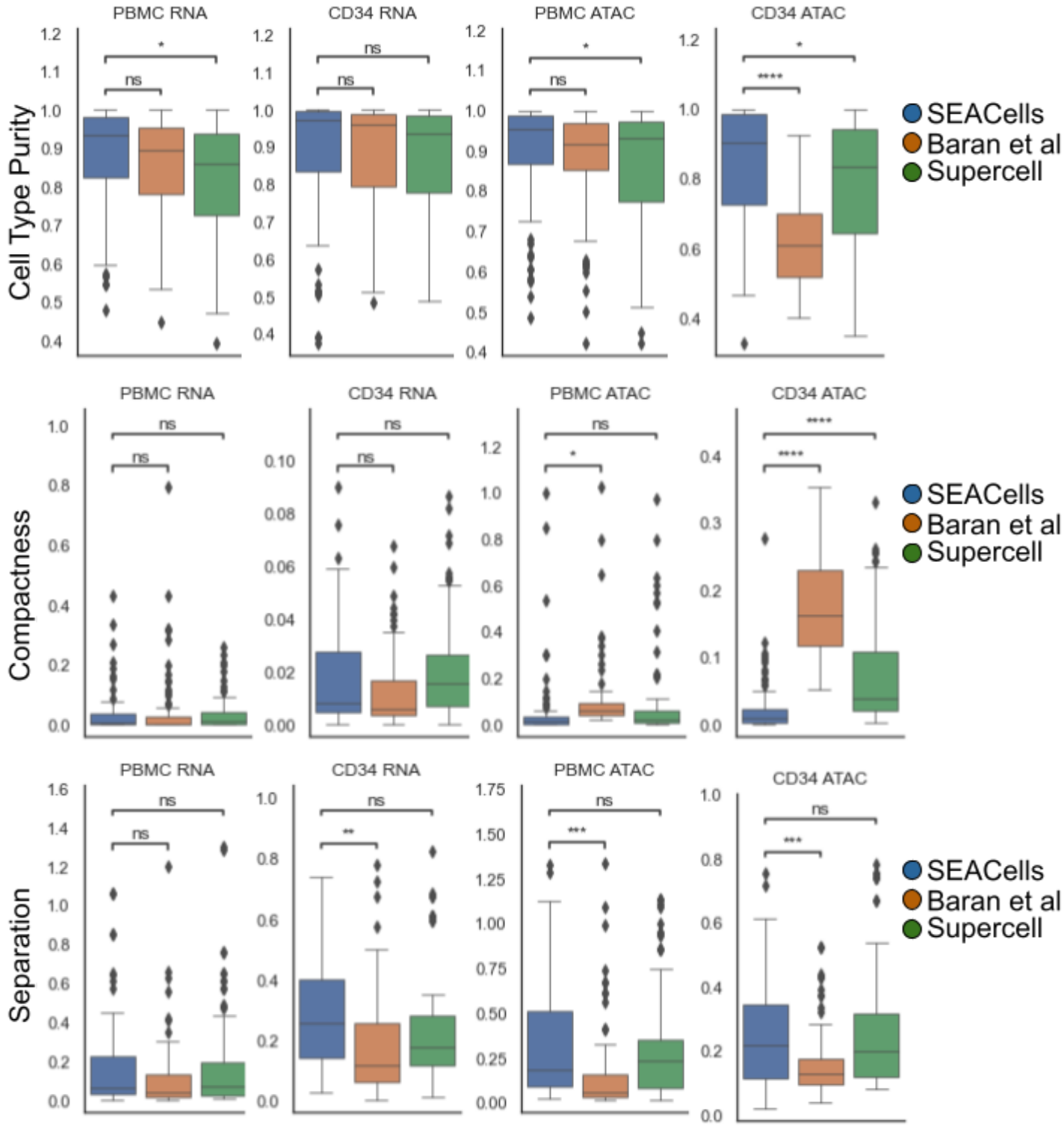


Figure A.1.

Appendix C: HERACLES Likelihood Function

Character Matrix

The input to our model is a character matrix, where each row represents a cell, and each column represents a target site. Denote by N the number of cells sequenced following an CRISPR-Cas9 lineage tracing experiment. Denote by σthe number of target sites, corresponding to each site in the recording cassette inserted into a given cell. The entries of this character matrix correspond to the state of each target site in a given cell - which may be unedited, mutated with a specific indel, or deleted. We denote the unedited state at site σ by ϕ_σ and index potential mutations at that site by $1, 2, \cdots, M_\sigma$. If a site is deleted, we represent its mutational state by D.
Therefore, for a given cell c, we denote its mutation state by at site σ by $\sigma_c \in \{\phi_\sigma, 1, \cdots, M_\sigma, D\}$.

Likelihood Function

The accumulation of CRISPR-Cas9 induced mutations at a given site evolves as a time-homogeneous, continuous time Markov Chain. At a given site, transitions may occur from an unedited state to a mutated or deleted state, or from a mutated state to a deleted state, depending on a site-specific mutation rate, λ_M, as well as a site specific transition probability matrix P^σ. Therefore, the transition probability function, $P^\sigma(t),\ t \geq 0$, is a row-stochastic matrix that defines a time-homogeneous and time-irreversible model of mutation. For any mutation states a, b, the entry $P^\sigma{}_{ab}(t)$ is the conditional probability of observing state b at any site t time units after observing state a at that site. We denote by Q^σ the infinitesimal generator of P^σ.

We can compute the evolutionary distance between two cells i and j, given their respective vectors of mutation states at each site in the recording cassette. Given that the mutation model in this lineage tracing system is a time-irreversible model, the distance between cells i and j at a particular site is determined by the distance in transitioning from the ancestral state to both σ_i and σ_j respectively. We will make the further simplifying assumption that the ancestral state is equidistant to both of the cells. As such, the data likelihood for a given pair of cells (i, j) separated by a tree distance t (representing evolutionary time) is given by

$$\mathcal{L}_{i,j}(t) = \prod_{\sigma} \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a P_{a\sigma_i}(\tfrac{t}{2}) P_{a\sigma_j}(\tfrac{t}{2})$$

where π_a is a prior over ancestral states and $A(\sigma_i, \sigma_j)$ denotes the set of all possible ancestral states for cells (i, j) at site σ. We are rescued from excessive complexity by the biological constraints of the system, which lead to a very limited set of possible ancestral states at a given site for a pair of cells, consisting of the unedited state and the least common ancestor (LCA) of the two cells, as illustrated in the table below.

Case	$A(\sigma_i, \sigma_j)$
$\sigma_i = \sigma_j$	$\{\phi_\sigma, \sigma_i\}$
$\sigma_i \neq \sigma_j$	$\{\phi_\sigma\}$

Therefore, over all pairs of cells, the log-likelihood of the data is given by

$$\log \mathcal{L}_{i,j}(t) = \sum_{\sigma} \log \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a P_{a\sigma_i}(\tfrac{t}{2}) P_{a\sigma_j}(\tfrac{t}{2})$$

The distance between cells i and j in hyperbolic space, denoted d_{ij}, serves as a proxy for tree distance, so our overall likelihood as a function of distances between points is therefore:

$$\log \mathcal{L}(d) = \sum_{i \neq j} \sum_{\sigma} \log \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a P_{a\sigma_i}(\tfrac{d_{ij}}{2}) P_{a\sigma_j}(\tfrac{d_{ij}}{2})$$

Transition Probability Matrix and Infinitesimal Generator

Inference under this model depends on the parameters of the transition probability matrix, P . These parameters can be informed by empirical knowledge of the behaviour of the system as determined by sensor screens performed *a priori* or can be inferred directly from observed data.

For a given site, the transition probability matrix has a very defined structure imposed by the biological assumptions of the CRISPR system, and is characterised by the following terms. For each site, the time taken to leave the base state and transition to either a mutated or deleted state follows an exponential distribution with rate $\lambda = \lambda_M + \lambda_D$, where λ_M corresponds to transitions to mutated stated and λ_D is a small value corresponding to transitions to a deleted state. Given that a transition occurs to some other character state, the transition to the observed character state, C, occurs with probability $p_{\phi_\sigma C}$.

Once a transition to a mutated state has occurred, there can be no further transitions, except in the case of a deletion, which occurs at some low rate of λ_D . As deletions are expected to be rarely observed, we will assume that all states and all sites have equal rates of transition to a deletion. Each site is, however, likely to have different mutation rates, λ_M as well as indel distributions (transition probabilities from base states to different characters).

Consider a toy lineage tracing system with three potential indel characters at two sites. The infinitesimal generator, Q, for the continuous time Markov Chain has the following structure, where $\sum_{i \in \{1,2,3\}} P_{\phi i} = 1$.

	ϕ_1	ϕ_2	1	2	3	D
ϕ_1	$-(\lambda_{M1} + \lambda_D)$	0	$\lambda_{M1} P_{\phi_1 1}$	$\lambda_{M1} P_{\phi_1 2}$	$\lambda_{M1} P_{\phi_1 3}$	λ_D
ϕ_2	0	$-(\lambda_{M2} + \lambda_D)$	$\lambda_{M2} P_{\phi_2 1}$	$\lambda_{M2} P_{\phi_2 2}$	$\lambda_{M2} P_{\phi_2 3}$	λ_D
1	0	0	$-\lambda_D$	0	0	λ_D
2	0	0	0	$-\lambda_D$	0	λ_D
3	0	0	0	0	$-\lambda_D$	λ_D
D	0	0	0	0	0	0

As several computations require computing the matrix exponential to determine transition probabilities, it will be helpful to express this matrix in a more compact form to take advantage of the structure of the transitions:

	ϕ_1	ϕ_2	M	D
ϕ_1	$-(\lambda_{M1} + \lambda_D)$	0	λ_{M1}	λ_D
ϕ_2	0	$-(\lambda_{M2} + \lambda_D)$	λ_{M2}	λ_D
M	0	0	$-\lambda_D$	λ_D
D	0	0	0	0

where M refers to any mutated character.

Estimating Mutation Rates and Indel Distributions

Given a tree topology, the transition probability matrix $P_{\phi_\sigma M}$ for a particular guide/site can be inferred by counting the number of branches along which such a transition occurred, and normalising by the number of total branches. The mutation rate of that site/guide can be computed as the total number of branches on which any transition to non-zero state occurred, normalised by the total number of branches. As such, we can use an initial tree to estimate the parameters of our system. Then, in an iterative manner, we can use optimization in hyperbolic space to optimise the topology given estimated parameters and use the improved tree to recompute updated model parameters.